Categorical Predictors I

2017-03-09

Preparation

We will use the data in the *mnSchools.csv* file. These data include institutional-level attributes for several Minnesota colleges and universities. The source of these data is: http://www.collegeresults.org. The attributes include:

- `id`: Institution ID number
- `name`: Institution name
- `gradRate`: Six-year graduation rate. This measure represents the proportion of first-time, full-time, bachelor's or equivalent degree-seeking students who started in Fall 2005 and graduated within 6 years.
- `public`: Dummy variable indicating educational sector (0 = private institution; 1 = public institution)
- `sat`: Estimated median SAT score for incoming freshmen at the institution
- `tuition`: Cost of attendance for full-time, first-time degree/certificate-seeking in-state undergraduate students living on campus for academic year 2013-14.

```
# Read in data
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 packages
library(dplyr)
library(ggplot2)
library(sm)
```

Exploration

Initially, we will plot the data. Note: Since the x-variable, public, is dummy coded, we need to turn it into a factor using `as.factor()` to get `ggplot()` to plot this correctly.

```
ggplot(data = mn, aes(x = as.factor(public), y = gradRate)) +
  geom_point() +
  theme_bw() +
  scale_x_discrete(name = "Educational sector", labels = c("Private", "Public")) +
  ylab("Six-year graduation rate")
```

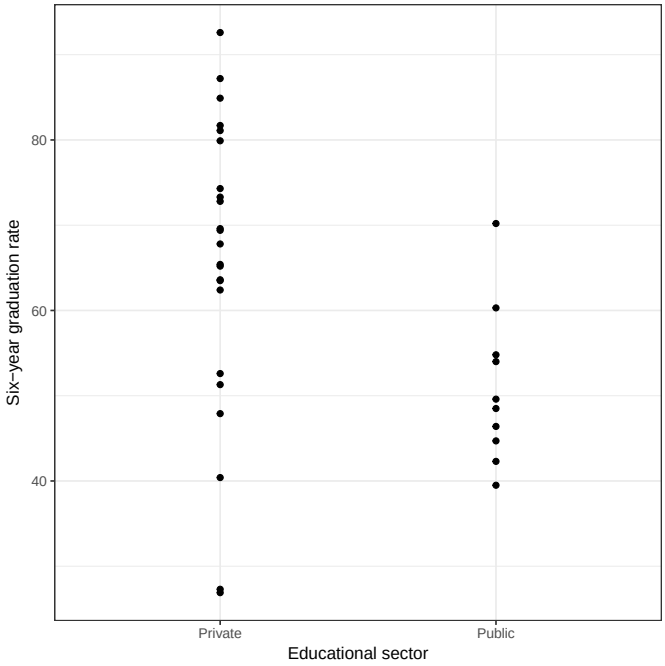

Figure 1. Scatterplot of the six-year graduation rate versus educational sector for $n = 33$ Minnesota colleges and universities.

Now, we will use the **dplyr** package to compute the means, standard deviations, and sample sizes for private (public = 0) and public (public = 1) schools.

```
mn %>%
  group_by(public) %>%
  summarize(
  M = mean(gradRate),
  SD = sd(gradRate),
  N = length(gradRate)
  )
```

Public	M	SD	N
0	65.27	17.58	23
1	51.03	9.16	10

We note a couple differences in the distribution of graduation rates between public and private schools. First, the mean graduation rates are different. Private schools have a graduation rate that is, on average, 14.2% higher than public schools. There is also more variation in private schools' graduation rates than in public schools' graduation rates. Lastly, we note that the sample sizes are not equal. There are 13 more private schools than there are public schools in the data set.

Simple Regression Model

Now we can fit the regression model to use educational sector (public/private) to predict variation in graduation rate.

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

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

Residuals:
    Min      1Q  Median      3Q     Max
-38.365  -6.330   0.135   9.035  27.335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   65.265      3.255  20.053   <2e-16 ***
public       -14.235      5.912  -2.408   0.0222 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.61 on 31 degrees of freedom
Multiple R-squared:  0.1575,    Adjusted R-squared:  0.1304
F-statistic: 5.797 on 1 and 31 DF,  p-value: 0.02219
```

Differences in sector explain 15.75% of the variation in graduation rates. This is statistically reliable, $F(1, 31) = 5.80$, $p = 0.022$. Interpreting the coefficients,

- The average graduation rate for private schools is 65.3%.
- Public schools, on average, have a graduation rate that is 14.2% lower than private schools.

The t-test associated with the slope coefficient suggests that the difference in means between private and public schools is likely different than 0 ($p = 0.022$). Given this evidence, we reject the hypothesis that $\beta_1 = 0$.

Re-Coding the Predictor

What happens if we had coded the predictor so that private schools were coded as 1, and public schools were coded as 0?

```
mn$private = ifelse(mn$public == 0, 1, 0)
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
  private
1       1
2       1
3       1
4       1
5       1
6       1
```

Now we use the `private` variable in the regression to predict variation in graduation rates.

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

```

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

Residuals:
    Min      1Q  Median      3Q     Max
-38.365  -6.330   0.135   9.035  27.335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.030      4.936  10.339 1.44e-11 ***
private       14.235      5.912   2.408   0.0222 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.61 on 31 degrees of freedom
Multiple R-squared:  0.1575,    Adjusted R-squared:  0.1304
F-statistic: 5.797 on 1 and 31 DF,  p-value: 0.02219
```

At the model-level, we end up with the same results. Differences in sector explain 15.75% of the variation in graduation rates. This is statistically reliable, $F(1, 31) = 5.80$, $p = 0.022$. Interpreting the coefficients,

- The average graduation rate for public schools is 51.0%.
- Private schools, on average, have a graduation rate that is 14.2% higher than public schools.

The results of the t-test associated with the slope coefficient is exactly the same as that where we used the `public` predictor, namely that there is likely a difference in means between private and public schools ($p = 0.022$). Given this evidence, we reject the hypothesis that $\beta_1 = 0$.

The only difference between the two fitted models is which sector's average graduation rate is expressed in the intercept. (The sign of the slope is also different.) This group is referred to as the *reference group*. In the first model we fitted, private schools were the reference group. In the second model, public schools were the reference group. The reference group will always be whichever group is coded as 0.

Assumption Checking

Like any other regression model, we need to examine whether or not the model's assumptions are satisfied. We look at (1) the marginal distribution of he standardized residuals, and (2) the scatterplot of the standardized residuals versus the model's fitted values.

```
# Use fortify() to obtain the fitted values and residuals
out = fortify(lm.public)
head(out)
```

```
  gradRate public       .hat   .sigma         .cooksd  .fitted
1     65.2      0 0.04347826 15.86645 0.0000004148202 65.26522
2     52.6      0 0.04347826 15.68931 0.0156443790139 65.26522
3     73.3      0 0.04347826 15.79540 0.0062962402814 65.26522
4     92.6      0 0.04347826 15.02351 0.0728726026129 65.26522
5     81.1      0 0.04347826 15.58867 0.0244544130063 65.26522
6     69.4      0 0.04347826 15.84767 0.0016673946113 65.26522
        .resid    .stdresid
1  -0.06521739 -0.004272246
2 -12.66521739 -0.829670222
3   8.03478261  0.526340738
4  27.33478261  1.790640811
5  15.83478261  1.037301389
6   4.13478261  0.270860412
```

Normality

```
# Density plot of the marginal standardized residuals
sm.density(out$.stdresid, model = "normal", xlab = "Studentized residuals")
```

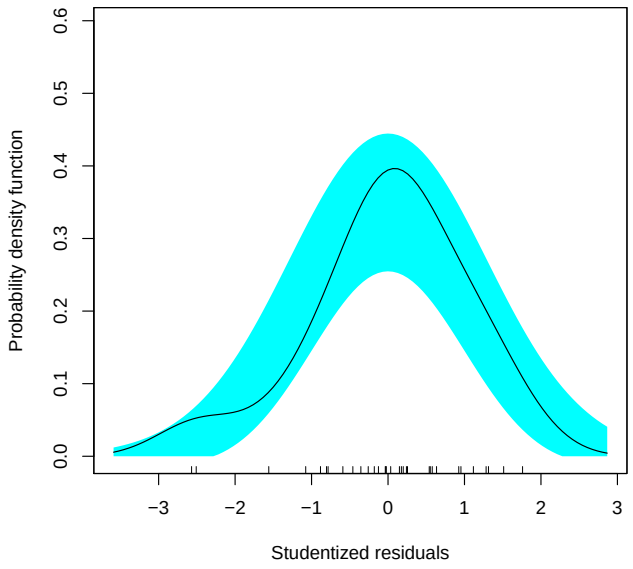

Figure 2. Density plot of the studentized residuals from the regression model using educational sector to predict variation in six-year graduation rates for $n = 33$ Minnesota colleges and universities.

The *marginal* distribution of the residuals does not show evidence of mis-fit with the normality assumption. Since the predictor has only two levels, we could actually examine the distribution of residuals for each sector. Here we do so as a pedagogical example, but note that once other non-categorical predictors are included, this can no longer be done.

Normality by Sector

We will use **dplyr** to filter the fortified data by sector.

```
out_private = out %>% filter(public == 0)
out_public = out %>% filter(public == 1)
```

Now we will plot each sector's residuals separately.

```
sm.density(out_private$.stdresid, model = "normal", xlab = "Studentized residuals")
```

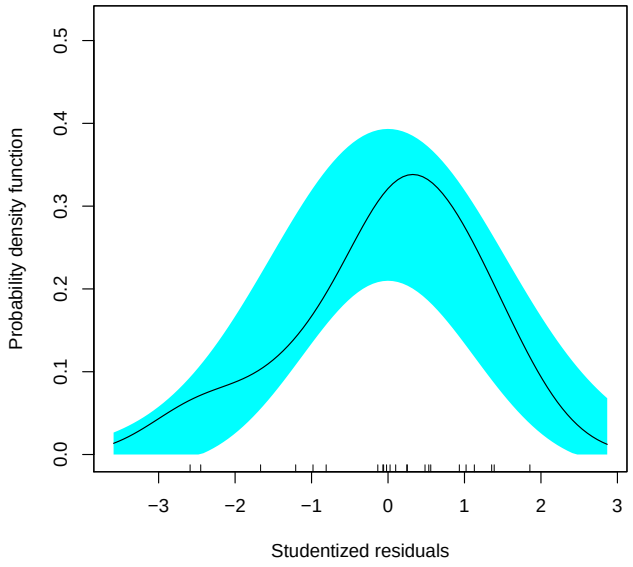

Figure 3. Density plot of the studentized residuals from the regression model using educational sector to predict variation in six-year graduation rates for $n = 23$ Minnesota private colleges and universities.

```
sm.density(out_public$.stdresid, model = "normal", xlab = "Studentized residuals")
```

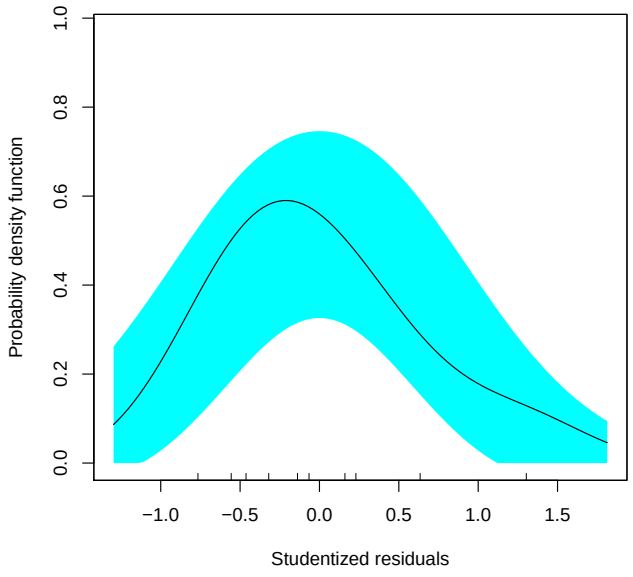

Figure 4. Density plot of the studentized residuals from the regression model using educational sector to predict variation in six-year graduation rates for $n = 10$ Minnesota public colleges and universities.

The normality assumption seems to be satisfied. Neither *conditional* distribution of residuals seem to indicate more mis-fit to normality than would be expected from sampling error.

Homoskedasticity

```
# Scatterplot of the standardized residuals versus the fitted values
ggplot(data = out, aes(x = .fitted, y = .stdresid)) +
  geom_point(size = 4) +
  theme_bw() +
  geom_hline(yintercept = 0) +
  xlab("Fitted values") +
  ylab("Studentized residuals")
```

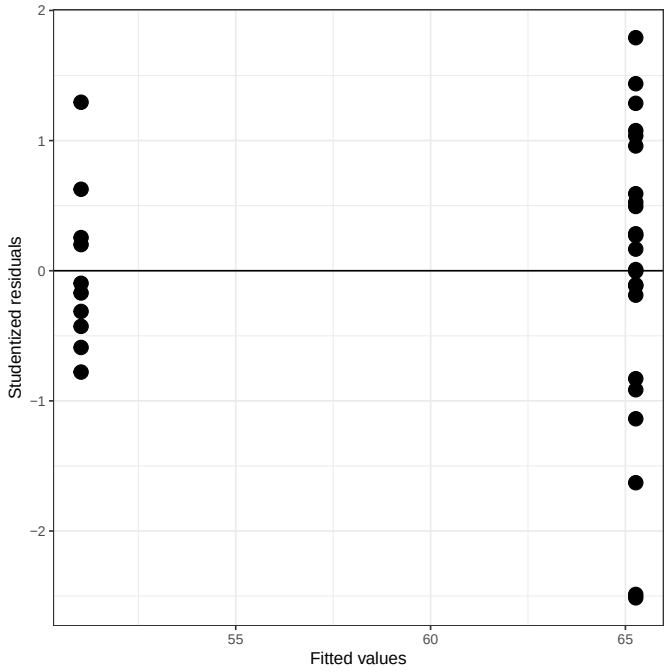

Figure 5. Scatterplot of the studentized residuals versus the fitted values from the regression model using educational sector to predict variation in six-year graduation rates for $n = 33$ Minnesota colleges and universities.

From this plot, we see that there is some question about the homoskedasticity assumption. We also saw that earlier when we examined the standard deviations of the two distributions. The variation in the private schools' residuals seems greater than the variation in the public schools' residuals. This, however, might be due to the private school that has a residual that is less than -2. This assumption violation might not be a problem once we add other predictors to the model, so for now, we will move on, but will re-check this assumption after fitting additional models.

Including Other Predictors

There seems to be differences between the average graduation rate between public and private institutions. It may be however, that the private schools are just more selective and this selectivity is the cause of the differences in graduation rates. To examine this, we will include the median SAT scores (`sat`) as a covariate into our model. So now, the regression model will include both the `public` dummy coded predictor and the `sat` predictors in an effort to explain variation in graduation rates.

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-19.9631  -3.1329   0.3607   4.9466  10.7336

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -76.05654   12.45216  -6.108 1.03e-06 ***
public       -8.37849    2.64822  -3.164  0.00355 **
sat           0.12672    0.01109  11.425 1.88e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.859 on 30 degrees of freedom
Multiple R-squared:  0.8426,    Adjusted R-squared:  0.8321
F-statistic: 80.27 on 2 and 30 DF,  p-value: 9.054e-13
```

Differences in sector explain 84.26% of the variation in graduation rates. This is statistically reliable, $F(2, 30) = 80.27$, $p < 0.001$. Interpreting the coefficients,

- The average graduation rate for private schools that have a 75th-percentile SAT score of 0 is -76.1%. (extrapolation)
- Public schools, on average, have a graduation rate that is 8.4% lower than private schools, contolling for differences in SAT scores.
- A ten-point differrence in SAT score is associated with a 1.3% difference in graduation rate, controlling for differences in sector.

The t-test associated with the slope coefficient for `public` suggests that the *controlled difference* in means between private and public schools is likely not 0 ($p = 0.004$). Given this evidence, we reject the hypothesis that $\beta_1 = 0$. This suggests that even after controlling for differences in SAT score, there is still a difference in private and public schools' graduation rates, on average.

Analysis of Covariance (ANCOVA)

Our research question fundamentally is: *Is there a difference on Y between group A and group B, after controlling for Z?* This is the simplest question stated in this form. We can make it more complex by having more than two groups (say group A, group B, and group C), or by controlling for multiple covariates. But, the primary question is whether there are group differences on some outcome.

In the social sciences, the methodology used to analyze group differences when controlling for other predictors is referred to as *analysis of covariance*, or ANCOVA. ANCOVA models can be analyzed using a framework

that focuses on partitioning variation (ANOVA) or using regression as a framework. Both ultimately give the same results (p-values, etc.). In this course we will focus using the regression framework to analyze this type of data.

Adjusted Means

Since the focus of the analysis is to answer whether there is a difference in graduation rates between private and public schools, we should provide some measure of how different the graduation rates are. Initially, we provided the mean graduation rates for public and private schools, along with the difference in these two means. These are referred to as the *unconditional means* and the *unconditional mean difference*, respectively. They are unconditional because they are the predicted means (y-hats) from the model that does not include any covariates.

After fitting our controlled model, we should provide new *adjusted means* and an *adjusted mean difference* based on the predicted mean graduation rates from the model that controls for SAT scores. Typically, the adjusted means are computed based on substituting in the mean value for all covariates, and then computing the predicted score for all groups. Here we show those computations for our analysis.

```
# Compute mean SAT
m_sat = mean(mn$sat)

# Compute adjusted means
d = expand.grid(
  public = c(0, 1),
  sat = m_sat
)

predict(lm.2, newdata = d)
```

```
       1        2
63.49045 55.11196
```

```
# Compute adjusted mean difference
63.5 - 55.1
```

```
[1] 8.4
```

Note that the adjusted mean difference is the value of the partial regression coefficient for `public` from the ANCOVA model. These values are typically presnted in a table along with the unadjusted values.

	Unadjusted Mean	Adjusted Mean
Private schools	65.3	63.5
Public schools	51.0	55.1
Difference	14.3	8.4

One Last Model

Now we will include the `public` dummy coded predictor, the `sat` predictor, and the `tuition` predictor in a model to explain variation in graduation rates. Our focus will be on whether or not there are mean differences in graduation rates between public and private schools, after controlling for differences in SAT scores and tuition.

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-17.3845  -2.9055   0.7195   3.9851  15.1959

Coefficients:
               Estimate  Std. Error t value    Pr(>|t|)
(Intercept) -68.2968905  12.5635462  -5.436 0.000007553 ***
public       -0.6468374   4.7155585  -0.137      0.8918
sat           0.1037931   0.0158652   6.542 0.000000364 ***
tuition       0.0004696   0.0002416   1.944      0.0617 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.562 on 29 degrees of freedom
Multiple R-squared:  0.8607,    Adjusted R-squared:  0.8463
F-statistic: 59.73 on 3 and 29 DF,  p-value: 1.587e-12
```

Differences in sector explain 86.07% of the variation in graduation rates. This is statistically reliable, $F(3, 29) = 59.73$, $p < 0.001$. Here we will not interpret all of the coefficients, but instead focus on only the `public` coefficient, as that is germaine to our research question.

- Public schools, on average, have a graduation rate that is 0.64% lower than private schools, contolling for differences in SAT scores and tuition.

The t-test associated with the partial slope coefficient for `public` suggests that the *controlled difference* in means between private and public schools is likely 0 ($p = 0.892$). Given this evidence, we fail to reject the hypothesis that $\beta_1 = 0$. This suggests that after controlling for differences in SAT score and tuition, there is not a difference in private and public schools' graduation rates, on average.

Assumption Checking for the Final Model

```r
# Use fortify() to obtain the fitted values and residuals
out3 = fortify(lm.3)
head(out3)
```

```
  gradRate public  sat tuition       .hat   .sigma     .cooksd   .fitted
1     65.2      0 1030   39294 0.10389064 6.477448 0.049739998  57.06260
2     52.6      0 1065   30480 0.08387593 6.632212 0.009081962  56.55627
3     73.3      0 1145   39400 0.04579828 6.627210 0.005278573  69.04858
4     92.6      0 1400   54265 0.26358723 6.312411 0.276390691 102.49648
5     81.1      0 1185   43198 0.06092474 6.570675 0.015004712  74.98386
6     69.4      0 1145   36590 0.05634885 6.670138 0.001025885  67.72899
     .resid  .stdresid
1  8.137400  1.3100117
2 -3.956266 -0.6299097
3  4.251419  0.6632600
4 -9.896481 -1.7574791
5  6.116139  0.9618275
6  1.671007  0.2621456
```

```r
# Density plot of the marginal standardized residuals
sm.density(out3$.stdresid, model = "normal")
```

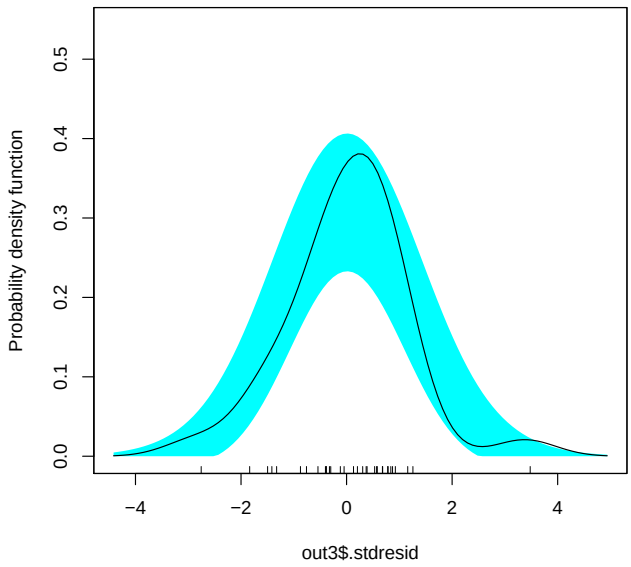

Figure 6. Density plot of the studentized residuals from the regression model using educational sector, median SAT, and tuition cost to predict variation in six-year graduation rates for $n = 33$ Minnesota colleges and universities.

```r
# Scatterplot of the standardized residuals versus the fitted values
ggplot(data = out3, aes(x = .fitted, y = .stdresid)) +
  geom_point(size = 4) +
  theme_bw() +
  geom_hline(yintercept = 0) +
  xlab("Fitted values") +
  ylab("Studentized residuals")
```

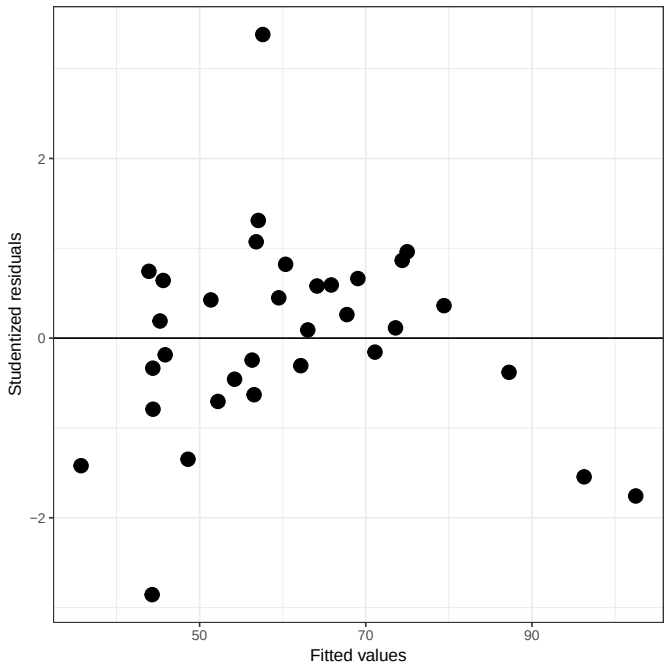

Figure 7. Scatterplot of the studentized residuals versus the fitted values from the regression model using educational sector, median SAT, and tuition cost to predict variation in six-year graduation rates for $n = 33$ Minnesota colleges and universities.

The marginal distribution of the residuals does not show evidence of mis-fit with the normality assumption. However, the scatterplot of the residuals versus the fitted values suggests clear problems with linaerity—at low fitted values more of the residuals are negative than we would expect (over-estimation); at moderate fitted values the residuals tend to be positive (under-estimation); and at high fitted values the residuals tend to the negative again (over-estimation). For now we will ignore this (although in practice this is a BIG problem).

Taxonomy of Models

Below we present pertinent results from the three models that we fitted.

Table 1. *Taxonomy of Models Examining the Effect of Educational Sector on Six-Year Graduation Rates for Minnesota Colleges and Universities ($n = 33$)*

	Model		
	(1)	(2)	(3)
Educational Sector	-14.235^{**}	-8.378^{***}	-0.647
	(5.912)	(2.648)	(4.716)
Median SAT score		0.127^{***}	0.104^{***}
		(0.011)	(0.016)
Tuition			0.0005^{*}
			(0.0002)
Constant	65.265^{***}	-76.057^{***}	-68.297^{***}
	(3.255)	(12.452)	(12.564)
R^2	0.158	0.843	0.861
RMSE	15.61	6.86	6.56

Note: *p<0.1; **p<0.05; ***p<0.01.
Educational Sector was dummy coded: 0 = Private; 1 = Public

Data Narrative

The presentation of the models help us build an evidence-based narrative about the differences in graduation rates between public and private schools. In the unconditional model, the evidence suggests that private schools have a higher graduation rate than public schools. Once we control for median SAT score, this difference in graduation rates persists, but at a much lesser magnitude. Finallly, after controlling for differences in SAT scores and tuition, we find no statistically reliable differences between the two educational sectors.

This narrative suggests that the initial differences we saw in graduation rates between the two sectors is really just a function of differences in SAT scores and tuition, and not really a public/private school difference. As with many non-experimental results, the answer to the question about group differences change as we control for different covariates. It may be, that once we control for other covariates, the narrative might change yet again. This is an important lesson, and one that cannot be emphasized enough—the magnitude and statistical importance of predictors change when the model is changed.