# Coding a Categorical Predictor

*2019-09-18*

## Preparation

In this set of notes, we will examine the question of whether there are differences in graduation rate between public and private colleges and universities. To do so, we will use the *mn-schools.csv* data (see the data codebook). To begin, we will load several libraries and import the data into an object called `mn`.

```
# Load packages
library(broom)
library(corrr)
library(dplyr)
library(educate)
library(ggplot2)
library(readr)
library(tidyr)

# Read in data
mn = read_csv(file = "~/Documents/github/epsy-8251/data/mn-schools.csv")
head(mn)
```

```
# A tibble: 6 x 5
  name                              grad public   sat tuition
  <chr>                            <dbl>  <dbl> <dbl>   <dbl>
1 Augsburg College                  65.2      0  1030    39.3
2 Bemidji State University          42.3      1  1010    18.1
3 Bethany Lutheran College          52.6      0  1065    30.5
4 Bethel University, Saint Paul, MN 73.3      0  1145    39.4
5 Carleton College                  92.6      0  1400    54.3
6 College of Saint Benedict         81.1      0  1185    43.2
```

## Exploration

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

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

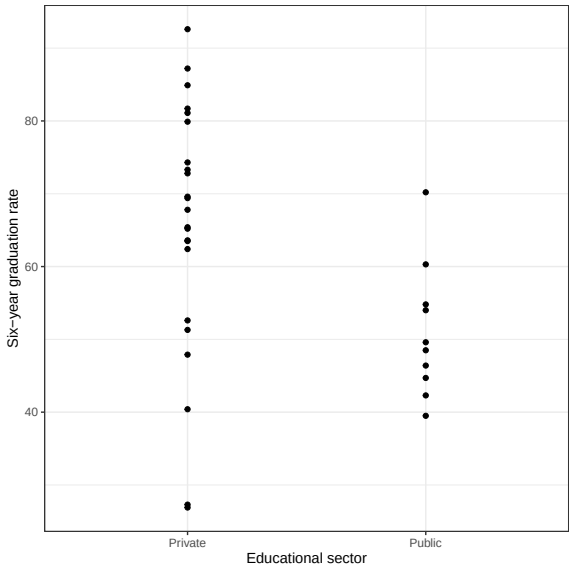



*Figure 1.* Scatterplot of the six-year graduation rate versus educational sector for $n = 33$ Minnesota colleges and universities.

We will also compute the means, standard deviations, and sample sizes for private (`public` = 0) and public (`public` = 1) schools.

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

Table 1
*Mean (M), Standard Deviation (SD), and Sample Size (N) of the Six-Year Graduation Rates for Private and Public Minnesota Colleges and Universities*

| Sector | M | SD | N |
|---|---|---|---|
| Private | 65.27 | 17.58 | 23 |
| Public | 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 percentage points 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.

Lastly, we will compute the pairwise correlation between educational sector and graduation rate.

```
mn %>%
  select(grad, public) %>%
  correlate() %>%
  fashion(decimals = 3)
```

```
  rowname  grad public
1    grad        -.397
2  public -.397
```

The correlation between educational sector and graduation rate is small and negative, indicating that institutions with higher graduation rates tend to have lower public values. Since there are only two values `public` can take, this implies that institutions with higher graduation rates tend to be private institutions; the lower value of `public` is 0 which corresponds to private institutions.

## 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(grad ~ 1 + public, data = mn)

glance(lm_public) #Model-level info
```

```
# 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     0.158         0.130  15.6      5.80  0.0222     2  -136.  279.  283.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

```
tidy(lm_public)   #Coefficient-level info
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     65.3      3.25     20.1  2.61e-19
2 public         -14.2      5.91     -2.41 2.22e- 2
```

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 percentage points 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 believe that the average graduation rate between private and public schools are, indeed, different.

## Reverse 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
```

```
# A tibble: 33 x 5
    grad public private   sat tuition
   <dbl>  <dbl>   <dbl> <dbl>   <dbl>
 1  65.2      0       1  1030    39.3
 2  42.3      1       0  1010    18.1
 3  52.6      0       1  1065    30.5
 4  73.3      0       1  1145    39.4
 5  92.6      0       1  1400    54.3
 6  81.1      0       1  1185    43.2
 7  69.4      0       1  1145    36.6
 8  47.9      0       1   990    37.8
 9  26.9      0       1   970    25.3
10  51.3      0       1  1030    33.2
# ... with 23 more rows
```

Now we use the `private` variable in the regression to predict variation in graduation rates. The results from fitting this regression model are shown below.

```
# 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     0.158         0.130  15.6      5.80  0.0222     2  -136.  279.  283.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     51.0      4.94     10.3  1.44e-11
2 private         14.2      5.91      2.41 2.22e- 2
```

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 percentage points 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 mean graduation rate between private and public schools ($p = 0.022$).

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 the studentized residuals, and (2) the scatterplot of the studentized residuals versus the model's fitted values. The only difference is that *with only categorical predictors in the model, we do not have to worry about the linearity assumption* since there is no ordinal quality to the predictor space.

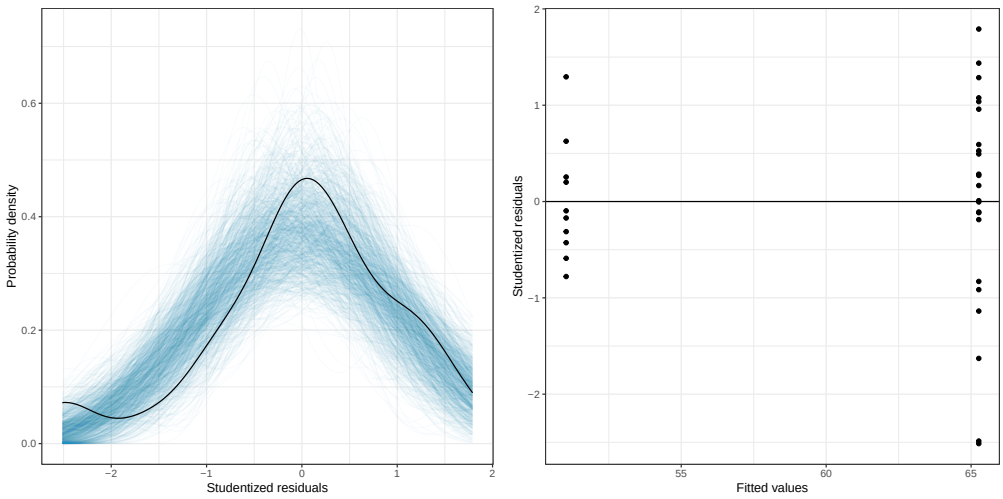
```
# Use augment() to obtain the fitted values and residuals
model_output = augment(lm_public)
head(model_output)
```

```
# A tibble: 6 x 9
   grad public .fitted .se.fit   .resid   .hat .sigma    .cooksd .std.resid
  <dbl>  <dbl>   <dbl>   <dbl>    <dbl>  <dbl>  <dbl>      <dbl>      <dbl>
1  65.2      0    65.3    3.25  -0.0652 0.0435   15.9    4.15e-7   -0.00427
2  42.3      1    51.0    4.94  -8.73   0.1      15.8    1.93e-2   -0.590
3  52.6      0    65.3    3.25 -12.7    0.0435   15.7    1.56e-2   -0.830
4  73.3      0    65.3    3.25   8.03   0.0435   15.8    6.30e-3    0.526
5  92.6      0    65.3    3.25  27.3    0.0435   15.0    7.29e-2    1.79
6  81.1      0    65.3    3.25  15.8    0.0435   15.6    2.45e-2    1.04
```

### Normality and Homoskedasticity

```
# Density plot of the marginal studentized residuals
ggplot(data = model_output, aes(x = .std.resid)) +
  stat_watercolor_density(model = "normal") +
  stat_density(geom = "line") +
  theme_bw() +
  xlab("Studentized residuals") +
  ylab("Probability density")

# Scatterplot of the studentized residuals versus the fitted values
ggplot(data = model_output, aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw() +
  xlab("Fitted values") +
  ylab("Studentized residuals")
```




From the scatterplot, we see that there is some question about the tenability of the homoskedasticity assumption. The variation in the private schools' residuals seems greater than the variation in the public schools' residuals. (This was foreshadowed earlier when we examined the standard deviations of the two distributions.) This difference in variation, however, might be due to the extreme private school that has a residual that is less than $-2$. Additionally, 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.

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

To examine the conditional distributions, we will filter the augmented data by sector and then plot each sector's residuals separately.

```
# Get private schools
model_output_private = model_output %>%
  filter(public == 0)

# Get public schools
model_output_public = model_output %>%
  filter(public == 1)

# Density plot of the private schools' sttudentized residuals
ggplot(data = model_output_private, aes(x = .std.resid)) +
  stat_watercolor_density(model = "normal") +
  stat_density(geom = "line") +
  theme_bw() +
  xlab("Studentized residuals") +
  ylab("Probability density")

# Density plot of the public schools' sttudentized residuals
ggplot(data = model_output_public, aes(x = .std.resid)) +
  stat_watercolor_density(model = "normal") +
  stat_density(geom = "line") +
```

```
  theme_bw() +
  xlab("Studentized residuals") +
  ylab("Probability density")
```

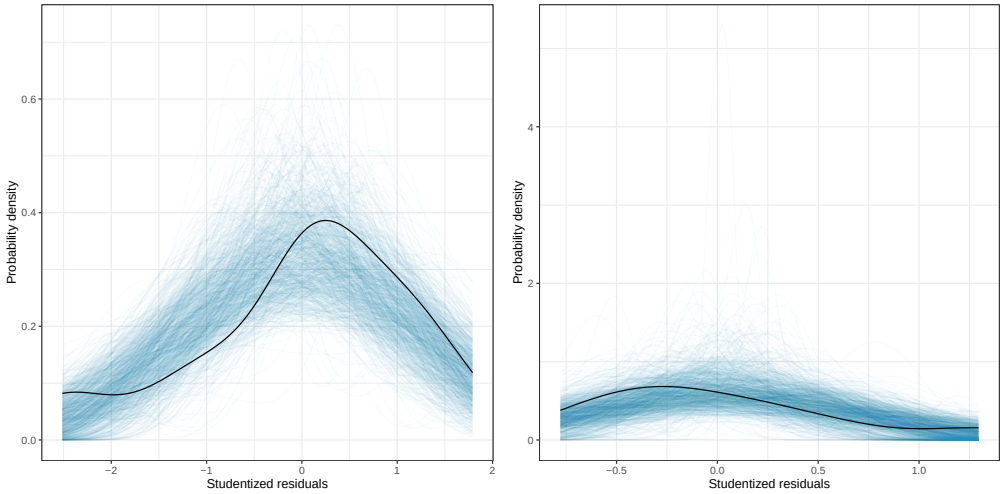



*Figure 2.* Density plot of the studentized residuals from the regression model using educational sector to predict variation in six-year graduation rates for Minnesota private (left) and public (right) colleges and universities. The bootstrapped confidence envelope (in blue) shows the sampling variation in density expected under the normality assumption. The scatterplot of the studentized residuals versus the fitted values from the same regression model is also displayed.

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

## 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.

Prior to fitting the regression model, we will examine the correlation matrix.

```
mn %>%
  select(grad, public, sat) %>%
  correlate() %>%
  fashion(decimals = 3)
```

```
  rowname  grad public   sat
1    grad        -.397  .889
2  public -.397        -.194
3     sat  .889  -.194
```

From the correlation matrix we see:

- Private institutions tend to have higher graduation rates than public institutions ($r = -0.397$).
- Institutions with higher median SAT scores tend to have higher graduation rates ($r = 0.889$).
- Private institutions tend to have higher median SAT scores than public institutions ($r = -0.397$).

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

# Model-level info
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     0.843         0.832  6.86      80.3 9.05e-13     3  -109.  226.  232.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

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

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

```
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -76.1     12.5        -6.11 1.03e- 6
2 public        -8.38     2.65       -3.16 3.55e- 3
3 sat            0.127    0.0111     11.4  1.88e-12
```

Interpreting the coefficients,

- The average graduation rate for private schools that have a median SAT score of 0 is $-76.1\%$. (extrapolation)
- Public schools, on average, have a graduation rate that is 8.4 percentage points lower than private schools, contolling for differences in median SAT scores.
- A ten-point differrence in median SAT score is associated with a 1.3 percentage point 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, even after controlling for differences in SAT score, we believe there is still a difference in private and public schools' graduation rates, on average.

## Analysis of Covariance (ANCOVA)

Our research question in the controlled model, fundamentally, is: *Is there a difference on Y between group A and group B, after controlling for covariate Z?* We can make the question 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 model 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
mean(mn$sat)
```

```
[1] 1101.212
```

```
# Compute adjusted mean for private schools
-76.1 - 8.4*0 + 0.127*1101.2
```

```
[1] 63.7524
```

```
# Compute adjusted mean for public schools
-76.1 - 8.4*1 + 0.127*1101.2
```

```
[1] 55.3524
```

```
# Compute adjusted mean difference
63.7 - 55.4
```

```
[1] 8.3
```

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.

Table 2
*Unadjusted and Adjusted Mean (Controlling for SAT Scores) Six-Year Graduation Rates for Private and Public Minnesota Colleges and Universities*

| | Unadjusted Mean | Adjusted Mean |
|---|---|---|
| Private institution | 65.3 | 63.5 |
| Public institution | 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.

Again, prior to fitting the regression model, we will examine the correlation matrix.

```
mn %>%
  select(grad, public, sat, tuition) %>%
  correlate() %>%
  fashion(decimals = 3)
```

```
  rowname  grad public   sat tuition
1    grad        -.397  .889    .755
2  public -.397        -.194   -.773
3     sat  .889  -.194          .613
4 tuition  .755  -.773  .613
```

From the correlation matrix we see:

- Private institutions tend to have higher graduation rates than public institutions ($r = -0.397$).
- Institutions with higher median SAT scores tend to have higher graduation rates ($r = 0.889$).
- Institutions with higher tuition costs tend to have higher graduation rates ($r = 0.755$).
- Private institutions tend to have higher median SAT scores than public institutions ($r = -0.397$).
- Private institutions tend to have higher tuition costs than public institutions ($r = -0.773$).
- Institutions with higher tuition costs tend to have higher median SAT scores ($r = 0.613$).

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

# Model-level info
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     0.861         0.846  6.56      59.7 1.59e-12     4  -107.  224.  231.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Differences in educational sector, median SAT scores, and tuition explain 86.07% of the variation in graduation rates. This is statistically reliable, $F(3, 29) = 59.73$, $p < 0.001$.

```
# Coefficient-level info
tidy(lm.3)
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic     p.value
  <chr>          <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)  -68.3     12.6       -5.44  0.00000755
2 public        -0.647    4.72      -0.137 0.892
3 sat            0.104    0.0159     6.54  0.000000364
4 tuition        0.470    0.242      1.94  0.0617
```

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 percentage points lower than private schools, contolling for differences in median SAT scores and tuition.

The $t$-test associated with the partial slope coefficient for `public` suggests that 0 is a possible value for the *controlled difference* in means between private and public schools ($p = 0.892$). Given this evidence, after controlling for differences in SAT score and tuition, it is uncertain thatthere is a difference in graduation rates between private and public schools, on average.

## Assumption Checking for the Final Model

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

```
# A tibble: 6 x 11
   grad public   sat tuition .fitted .se.fit .resid   .hat .sigma .cooksd
  <dbl>  <dbl> <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl>   <dbl>
1  65.2      0  1030    39.3    57.1    2.12   8.14 0.104    6.48 0.0497
2  42.3      1  1010    18.1    44.4    2.21  -2.07 0.113    6.67 0.00357
3  52.6      0  1065    30.5    56.6    1.90  -3.96 0.0839   6.63 0.00908
4  73.3      0  1145    39.4    69.0    1.40   4.25 0.0458   6.63 0.00528
5  92.6      0  1400    54.3   102.     3.37  -9.90 0.264    6.31 0.276
6  81.1      0  1185    43.2    75.0    1.62   6.12 0.0609   6.57 0.0150
# ... with 1 more variable: .std.resid <dbl>
```

```
# Density plot of the marginal studentized residuals
ggplot(data = lm.3_output, aes(x = .std.resid)) +
  stat_watercolor_density(model = "normal") +
  stat_density(geom = "line") +
  theme_bw() +
  xlab("Studentized residuals") +
  ylab("Probability density")

# Scatterplot of the studentized residuals versus the fitted values
ggplot(data = lm.3_output, aes(x = .fitted, y = .std.resid)) +
  geom_point(size = 4) +
  geom_smooth(se = TRUE) +
  geom_hline(yintercept = 0) +
  theme_bw() +
  xlab("Fitted values") +
  ylab("Standardized residuals")
```

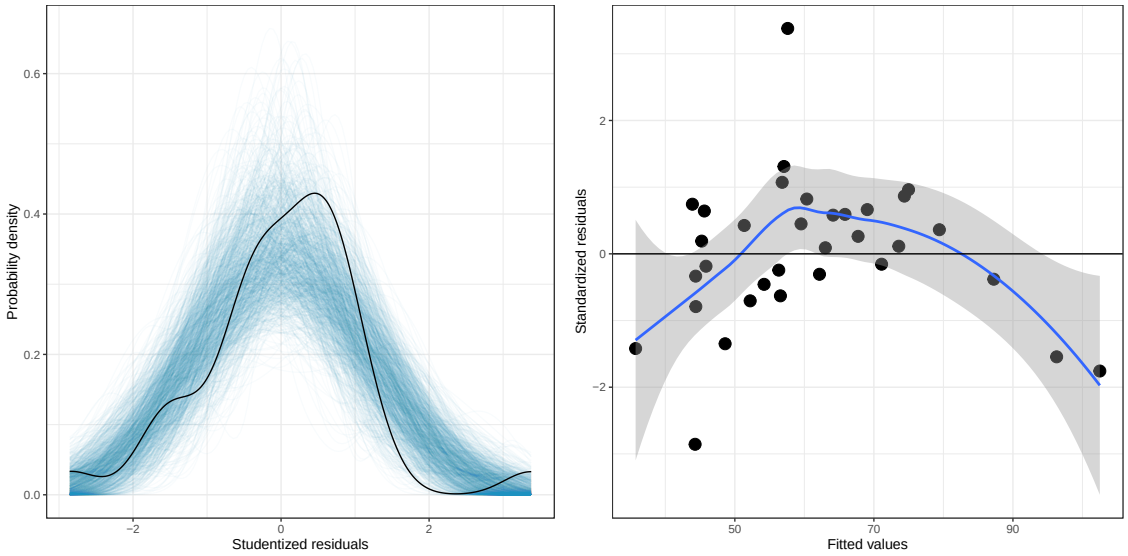



*Figure 3*. Density plot of the studentized residuals from the regression model using educational sector, median SAT, and tuition cost to predict variation in six-year graduation rates for $n = 33$ Minnesota colleges and universities. The bootstrapped confidence envelope (in blue) shows the sampling variation in density expected under the normality assumption. The scatterplot of the sttudentized residuals versus the fitted values from the same regression model is also shown.

The marginal distribution of the residuals shows minor evidence of mis-fit with the normality assumption. Homoskedasticity looks tenable from this plot. However, the scatterplot of the residuals versus the fitted values suggests problems with the tenability of the linearity assumption—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.

### 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.

Table 3
*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) |
| Public institution† | −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.
†Public institution was dummy coded: 0 = Private; 1 = Public