# Categorical Predictors I

*2018-07-27*

## 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.

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

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

```
# A tibble: 6 x 6
     id name                              gradRate public   sat tuition
  <int> <chr>                                <dbl>  <int> <int>   <int>
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
```

## 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.

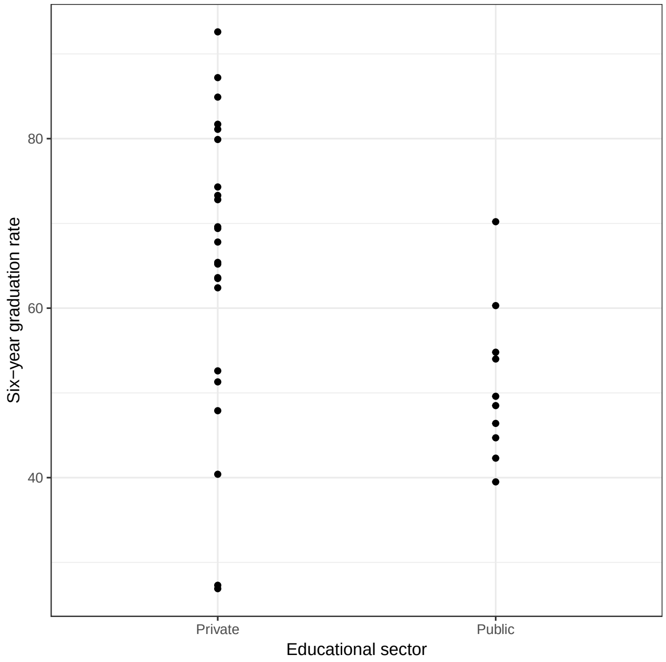


Figure 1: Scatterplot of the six-year graduation rate versus educational sector for $n = 33$ Minnesota colleges and universities.

```r
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")
```

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.

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

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% 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(gradRate, public) %>%
  correlate() %>%
  fashion(decimals = 3)
```

```
   rowname gradRate public
1 gradRate            -.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(gradRate ~ 1 + public, data = mn)

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

```
  r.squared adj.r.squared    sigma statistic    p.value df    logLik
1 0.1575459       0.13037 15.60845  5.797258 0.02219125  2 -136.4712
       AIC      BIC deviance df.residual
1 278.9424 283.4319 7552.333          31
```

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

```
         term   estimate std.error statistic      p.value
1 (Intercept)  65.26522  3.254586  20.05331 2.608064e-19
2      public -14.23522  5.912250  -2.40775 2.219125e-02
```

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 $\beta_1 = 0$.

## 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
   gradRate public private   sat tuition
      <dbl>  <int>   <dbl> <int>   <int>
 1     65.2      0       1  1030   39294
 2     52.6      0       1  1065   30480
 3     73.3      0       1  1145   39400
 4     92.6      0       1  1400   54265
 5     81.1      0       1  1185   43198
 6     69.4      0       1  1145   36590
 7     47.9      0       1   990   37795
 8     26.9      0       1   970   25345
 9     51.3      0       1  1030   33210
10     81.7      0       1  1225   43800
# ... 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.

```
  r.squared adj.r.squared    sigma statistic    p.value df    logLik
1 0.1575459       0.13037 15.60845  5.797258 0.02219125  2 -136.4712
       AIC      BIC deviance df.residual
1 278.9424 283.4319 7552.333          31
```

```
         term estimate std.error statistic      p.value
1 (Intercept) 51.03000  4.935825  10.33870 1.437896e-11
2     private 14.23522  5.912250   2.40775 2.219125e-02
```

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

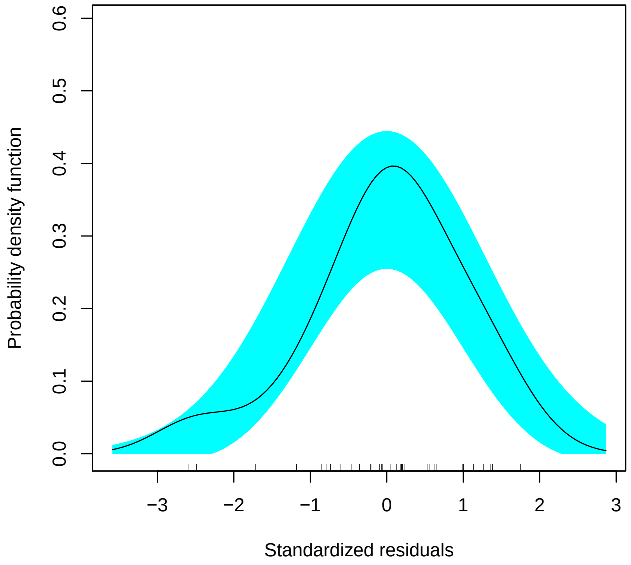
Like any other regression model, we need to examine whether or not the model's assumptions are satisfied. We look at (1) the marginal distribution of he standardized residuals, and (2) the scatterplot of the standardized residuals versus the model's fitted values.

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

```
  gradRate public  .fitted  .se.fit        .resid        .hat   .sigma
1     65.2      0 65.26522 3.254586  -0.06521739 0.04347826 15.86645
2     52.6      0 65.26522 3.254586 -12.66521739 0.04347826 15.68931
3     73.3      0 65.26522 3.254586   8.03478261 0.04347826 15.79540
4     92.6      0 65.26522 3.254586  27.33478261 0.04347826 15.02351
5     81.1      0 65.26522 3.254586  15.83478261 0.04347826 15.58867
6     69.4      0 65.26522 3.254586   4.13478261 0.04347826 15.84767
          .cooksd   .std.resid
1 0.0000004148202 -0.004272246
2 0.0156443790139 -0.829670222
3 0.0062962402814  0.526340738
4 0.0728726026129  1.790640811
5 0.0244544130063  1.037301389
6 0.0016673946113  0.270860412
```

### Normality

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



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$.std.resid, model = "normal", xlab = "Standardized residuals")
sm.density(out_public$.std.resid, model = "normal", xlab = "Standardized residuals")
```

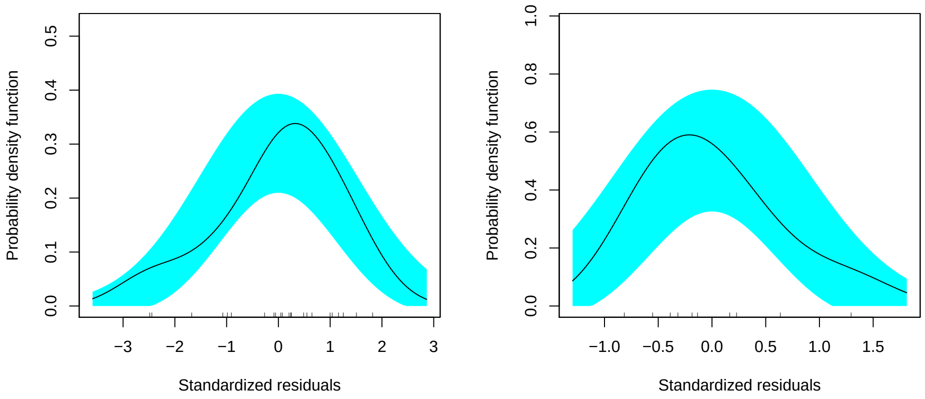


Figure 2: Density plot of the standardized residuals from the regression model using educational sector to predict variation in six-year graduation rates for Minnesota private (left) and public (right) colleges and universities.

The 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 = .std.resid)) +
  geom_point(size = 4) +
  theme_bw() +
  geom_hline(yintercept = 0) +
  xlab("Fitted values") +
  ylab("Standardized residuals")
```

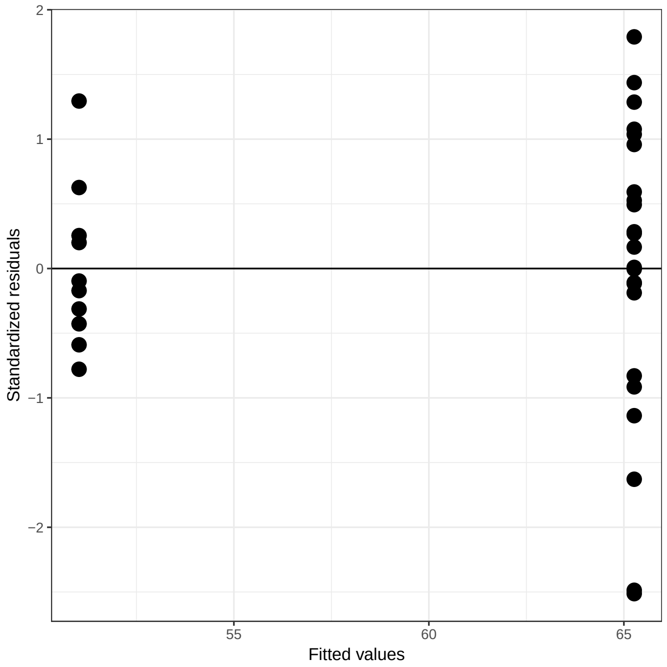


Figure 3: Scatterplot of the standardized residuals versus the fitted values from the regression model using educational sector to predict variation in six-year graduation rates for $n = 33$ Minnesota colleges and universities.

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.

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

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

```
   rowname gradRate public   sat
1 gradRate           -.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(gradRate ~ 1 + public + sat, data = mn)

# Model-level info
glance(lm.2)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1 0.8425573     0.8320611 6.859116  80.27274 9.053974e-13  3 -108.7964
       AIC      BIC deviance df.residual
1 225.5929 231.5789 1411.424          30
```

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)
```

```
         term    estimate   std.error statistic      p.value
1 (Intercept) -76.0565386 12.45215604 -6.107901 1.031640e-06
2      public  -8.3784912  2.64822241 -3.163817 3.554453e-03
3         sat   0.1267213  0.01109178 11.424789 1.884787e-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% 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% 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 null 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 in the controlled model, 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.

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

# Compute adjusted means
avg_inst = crossing(
  public = c(0, 1),
  sat = m_sat
)

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

```
       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.

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(gradRate, public, sat, tuition) %>%
  correlate() %>%
  fashion(decimals = 3)
```

```
   rowname gradRate public   sat tuition
1 gradRate           -.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(gradRate ~ 1 + public + sat + tuition, data = mn)

# Model-level info
glance(lm.3)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1  0.860709     0.8462996 6.561906  59.73241 1.587384e-12  4 -106.7753
       AIC     BIC deviance df.residual
1 223.5505 231.033   1248.7          29
```

Differences in sector 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)
```

```
         term       estimate     std.error  statistic        p.value
1 (Intercept) -68.2968905146 12.5635461864 -5.4361157 0.000007552786
2      public  -0.6468374221  4.7155584912 -0.1371709 0.891843634272
3         sat   0.1037930682  0.0158651540  6.5422037 0.000000364095
4     tuition   0.0004696043  0.0002415658  1.9440012 0.061654397872
```

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

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

```
  gradRate public  sat tuition   .fitted  .se.fit    .resid        .hat
1     65.2      0 1030   39294  57.06260 2.115038  8.137400 0.10389064
2     52.6      0 1065   30480  56.55627 1.900416 -3.956266 0.08387593
3     73.3      0 1145   39400  69.04858 1.404283  4.251419 0.04579828
4     92.6      0 1400   54265 102.49648 3.368931 -9.896481 0.26358723
5     81.1      0 1185   43198  74.98386 1.619671  6.116139 0.06092474
6     69.4      0 1145   36590  67.72899 1.557659  1.671007 0.05634885
    .sigma     .cooksd .std.resid
1 6.477448 0.049739998  1.3100117
2 6.632212 0.009081962 -0.6299097
3 6.627210 0.005278573  0.6632600
4 6.312411 0.276390691 -1.7574791
5 6.570675 0.015004712  0.9618275
6 6.670138 0.001025885  0.2621456
```

```
# Density plot of the marginal standardized residuals
sm.density(out3$.std.resid, model = "normal", xlab = "Standardized residuals")
```

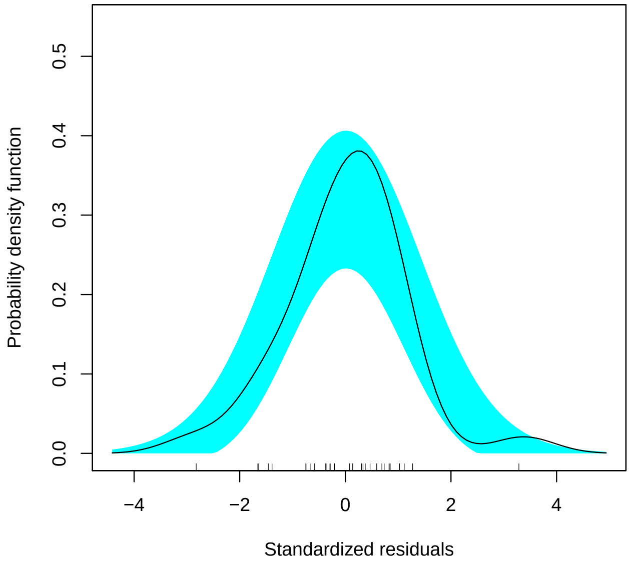


Figure 4: Density plot of the standardized residuals from the regression model using educational sector, median SAT, and tuition cost to predict variation in six-year graduation rates for $n = 33$ Minnesota colleges and universities.

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

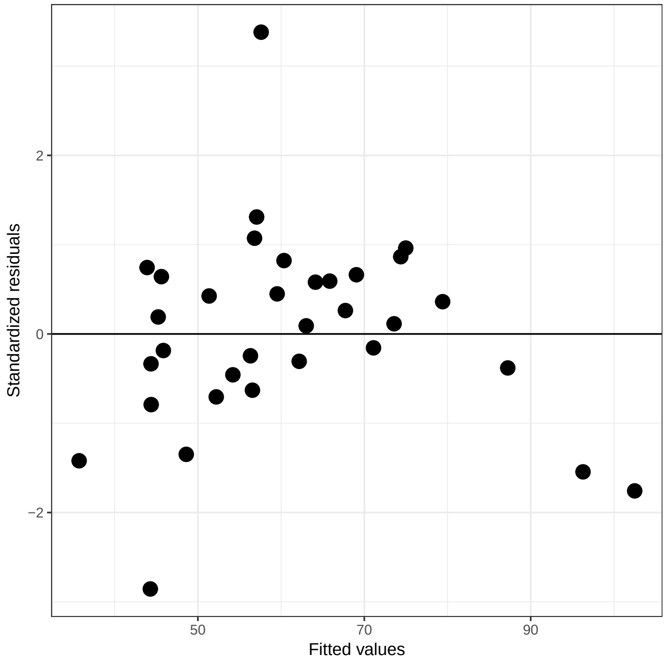


Figure 5: Scatterplot of the standardized residuals versus the fitted values from the regression model using educational sector, median SAT, and tuition cost to predict variation in six-year graduation rates for $n = 33$ Minnesota colleges and universities.

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 linearity—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 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 | $-14.235^{**}$<br>(5.912) | $-8.378^{***}$<br>(2.648) | $-0.647$<br>(4.716) |
| Median SAT score | | $0.127^{***}$<br>(0.011) | $0.104^{***}$<br>(0.016) |
| Tuition | | | $0.0005^{*}$<br>(0.0002) |
| Constant | $65.265^{***}$<br>(3.255) | $-76.057^{***}$<br>(12.452) | $-68.297^{***}$<br>(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

## 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.

## Plots to Display Model Results

There are two plots we may want to consider creating to accompany the data narrative: (1) a coefficient plot to emphasize the public–private institution difference (i.e., effect of sector), and (2) a plot of the final fitted model, again emphasizing the public–private institution difference.

### Coefficient Plot

```r
# Create tidy model objects
m1 = tidy(lm_public) %>% mutate(model = "Model 1")
m2 = tidy(lm.2) %>% mutate(model = "Model 2")
m3 = tidy(lm.3) %>% mutate(model = "Model 3")

# Bind into a single object
all_models = rbind(m1, m2, m3)

# Coefficient plot
dw_plot(all_models, show_intercept = FALSE) +
  theme_bw() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank()
  ) +
  geom_vline(xintercept = 0, linetype = "dashed", alpha = 0.4) +
  scale_y_discrete(
    name = "",
    labels = c("Tuition", "Median SAT score", "Public")
  ) +
  scale_color_manual(
    name = "",
    labels = c("Model 1", "Model 2", "Model 3"),
    values = c("#999999", "#e69f00", "#56b4e9")
  )
```

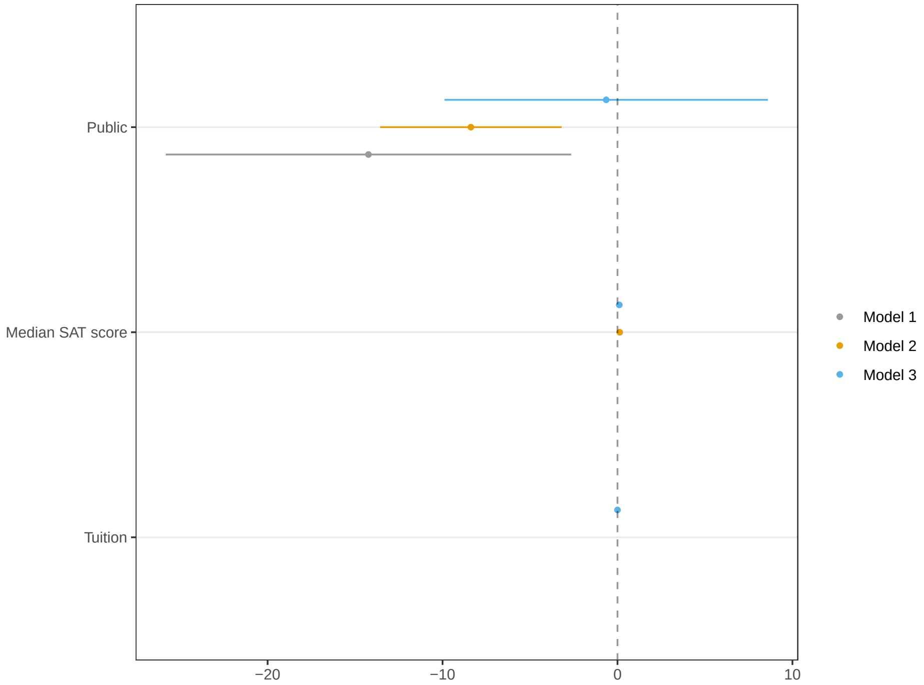


Figure 6: Plot displaying the fitted coefficient estimates and 95% confidence intervals for three models regressing graduation rate on educational sector (public) and two institutional covariates.

### Plot of the Final Fitted Model

```
# Set up the data to plot
plot_data = crossing(
  sat = seq(from = 890, to = 1400, by = 10),
  public = c(0, 1),
  tuition = mean(mn$tuition)
  )

plot_data %>%
  mutate(
    # Get predicted values
    yhat = predict(lm.3, newdata = plot_data),
    # Change public into a factor
    sector = factor(public,
                    levels = c(0, 1),
                    labels = c("Public institution", "Private institution")
                    )
    ) %>%
  # Create plot
  ggplot(aes(x = sat, y = yhat, group = sector, color = sector)) +
      geom_line() +
      theme_bw() +
      xlab("Median SAT score") +
      ylab("Predicted graduation rate") +
      scale_color_viridis_d(name = "")
```

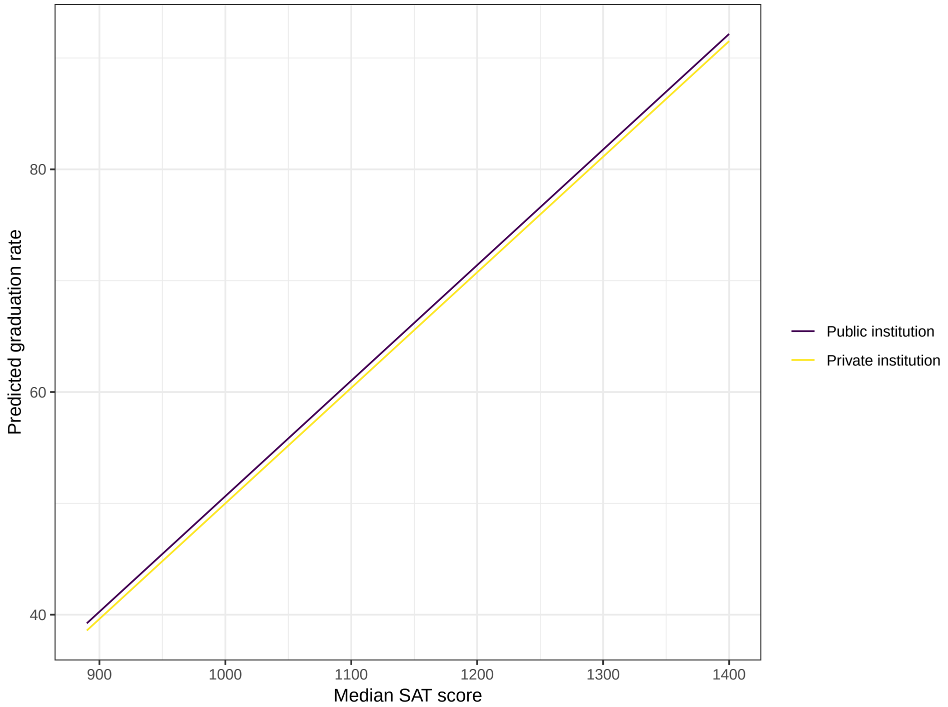


Figure 7: Predicted graduation rate as a function of median SAT scores for public and private institutions in Minnesota. Tuition rate is controlled for by fixing the value to the marginal average tuition value.

In my opinion, the coefficient plot corresponds better to the data narrative than the plot of the fitted model, so in a publication that is what I would present along with the table of model results.