More Categorical Predictors

2017-10-24

Preparation

In this set of notes, you will continue learning about the inclusion of categorical predictors in regression models. We will use data collected by fivethirtyeight to examine differences in the median incomes across different STEM majors. In particular, we will focus on whether women are attracted to STEM majors that have lower median incomes. In fivethirtyeight's analysis, they suggested that women are more attracted to the "S"-majors than the "TEM"-majors, and it is the "S"-majors that have lower median incomes. We will examine this via fitting a series of regression models. The dataset, *stem.csv*, includes data on 76 STEM majors, including:

- `major`: Name of STEM major
- `income`: Median income (in thousands of dollars) for a full-time, year-round worker
- `women`: Percentage of recent graduates who are women
- `stem_type`: Type of STEM major (Science; Technology, Engineering, Mathematics)

```
# Load libraries
library(broom)
library(corrr)
library(dplyr)
library(ggplot2)
library(sm)
library(readr)

# Read in data
stem = read_csv(file = "~/Dropbox/epsy-8251/data/stem.csv")
head(stem)
```

```
# A tibble: 6 x 4
                               major income    women   stem_type
                               <chr>  <dbl>    <dbl>       <chr>
1                Aerospace Engineering     60 13.97928 Engineering
2                  Applied Mathematics     45 43.42984 Mathematics
3            Architectural Engineering     54 35.04425 Engineering
4                        Architecture     40 45.14649 Engineering
5           Astronomy And Astrophysics     62 53.57143     Science
6 Atmospheric Sciences And Meteorology     35 32.12961     Science
```

Examine and Describe the Marginal Distribution of the Median Incomes

```
sm.density(stem$income, xlab = "Median income")
```

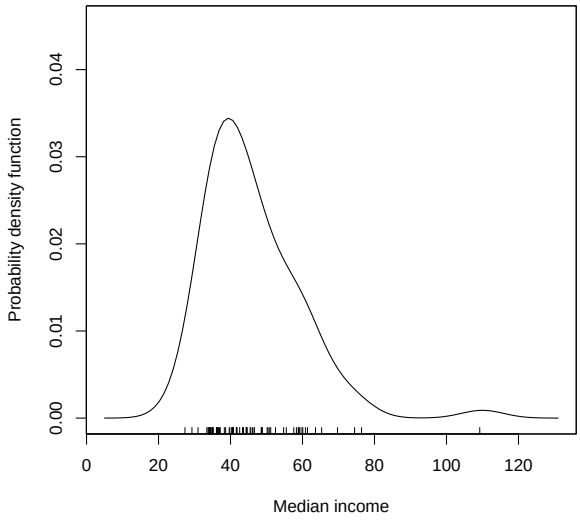

Figure 1. Density plot of the median incomes for $n = 76$ STEM majors.

```
# Compute summary statistics
stem %>%
  summarize(
    M  = mean(income),
    SD = sd(income),
    Min = min(income),
    Max = max(income)
  )
```

```
# A tibble: 1 x 4
         M       SD   Min   Max
     <dbl>    <dbl> <dbl> <dbl>
1 46.11842 13.18722    26   110
```

The median incomes for the 76 STEM majors are right-skewed and range from $26,000 to $110,000. The mean income is $46,000. The standard deviation is $13,187.

Are Women Attracted to Lower-Earning STEM Majors?

To answer this question, we can examine a scatterplot of the relationship between the percentage of recent graduates in each major that are women, and the median income for those majors.

```
# Plot the incomes by proportion of women
ggplot(data = stem, aes(x = women, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw() +
    xlab("Percentage of recent graduates who are women") +
    ylab("Median income (in thousands of dollars)")
```

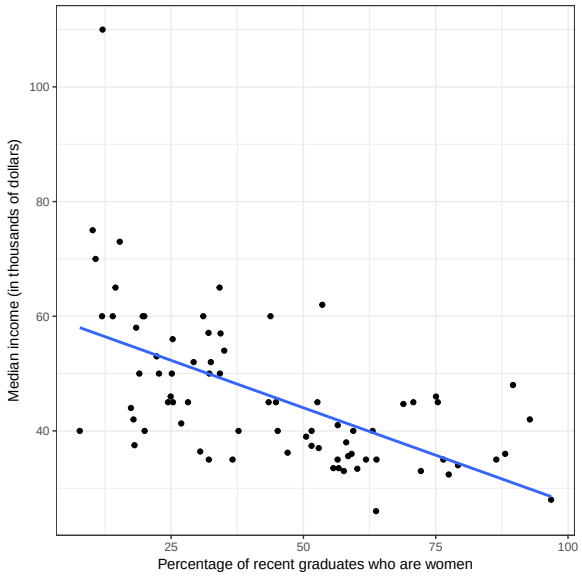

Figure 2. Scatterplot showing the relationship between the percentage of recent graduates in each STEM major that are women, and the median income for those majors. The OLS regression lines is also displayed on the plot.

```
# Compute correlation coefficient
stem %>%
  select(income, women) %>%
  correlate() %>%
  shave() %>%
  fashion(decimals = 3)
```

```
  rowname income women
1  income
2   women  -.583
```

The scatterplot suggests a negative relationship between these variables. This implies that majors that have a higher percentage of femal graduates tend to be the same majors that have lower median incomes ($r = -.583$). This relationship seems linear and moderately strong. There is one major (Petroleum Engineering) that has an unusually high median income ($110,000).

Fitting a Regression Model

```
lm.1 = lm(income ~ 1 + women, data = stem)
summary(lm.1)
```

```
Call:
lm(formula = income ~ 1 + women, data = stem)

Residuals:
    Min      1Q  Median      3Q     Max
-18.014  -6.471  -0.900   5.246  53.413

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept) 60.57646    2.65139  22.847      < 2e-16 ***
women       -0.33090    0.05366  -6.166 0.0000000337 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.79 on 74 degrees of freedom
Multiple R-squared:  0.3394,    Adjusted R-squared:  0.3305
F-statistic: 38.02 on 1 and 74 DF,  p-value: 0.00000003369
```

The fitted regression model is

$$\hat{\text{Median Income}} = 60.58 - 0.33(\text{Percentage of Women})$$

- The fitted intercept ($\hat{\beta}_0 = 60.58$) indicates that the median income for STEM majors that are 100% male is \$60,580, on average.
- The fitted slope ($\hat{\beta}_1 = -0.33$) indicates that, on average, the median income difference for STEM majors that have one-percent more female graduates is \$330 less.

The p-value associated with the slope suggests that this difference in income is statistically significant.

Are there Income Differences by the Type of STEM Major?

```
# Plot the median incomes by STEM type
ggplot(data = stem, aes(x = stem_type, y = income, fill = stem_type)) +
  geom_point(shape = 21, color = "black", size = 4) +
    theme_bw() +
    xlab("Type of STEM Major") +
    ylab("Median income (in thousands of dollars)") +
  guides(fill = FALSE)
```

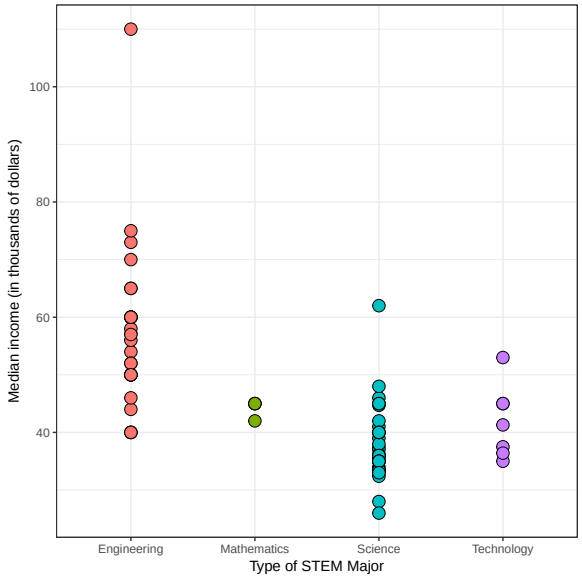

Figure 3. Scatterplot showing the relationship between the type of STEM major and the median income for those majors.

```
stem %>%
  group_by(stem_type) %>%
  summarize( M = mean(income), SD = sd(income), N = n() )
```

```
# A tibble: 4 x 4
    stem_type        M        SD     N
        <chr>    <dbl>     <dbl> <int>
1 Engineering 57.38276 13.626080    29
2 Mathematics 44.25000  1.500000     4
3     Science 38.07500  6.434433    36
4  Technology 41.88571  6.324668     7
```

The sample data suggests that there are potential income differences across the STEM major types. The mean income for "S"-majors is lower than for "TEM"-majors. In particular, "T"- and "E"-majors seem to have similar average incomes (around $43,000). These majors earn roughly $5,000 more than "S"-majors, but earn $14,000 less than "E"-majors. However, there is variation within each of the major types.

Ridge Plots: An Alternative PLot for Comparing Distributions

Ridge plots are partially overlapping density plots that create the impression of a mountain range. They can be useful for comparing distributions. For more information, see the package vignette.

```
# Load the ggjoy package
library(ggridges)

# Ridge plot
ggplot(data = stem, aes(x = income, y = stem_type)) +
  geom_density_ridges() +
  theme_bw() +
  xlab("Median income") +
  ylab("Type of STEM major")
```

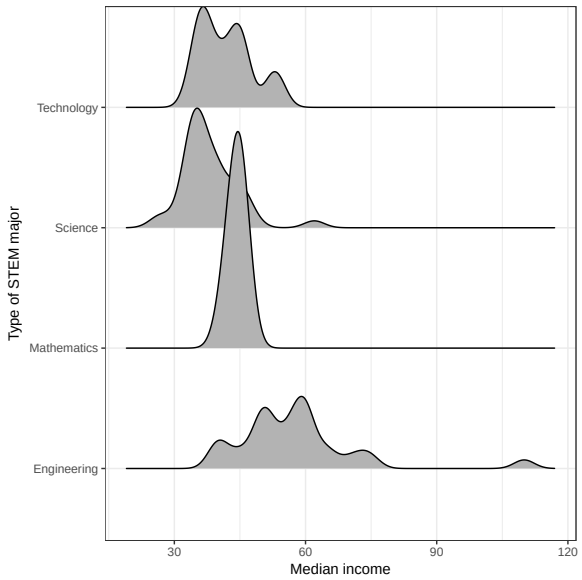

Figure 4. Ridge plot showing the distribution of median income for four types of STEM majors.

This plot suggests the same pattern in average incomes. The plots for the "S"-, "T"-, and "M"-majors all overlap and have a similar average, while the distribution for the incomes for the "E"-majors is located further to the right (higher average income).

Fitting a Regression Model

Before fitting a regression model, we need to create a dummy variable for EACH category of the `stem_type` variable. For our analysis, we will need to create four dummy variables: `science`, `tech`, `engineer`, and `math`. To do this we will use the `if_else()` function.

The `if_else()` function evaluates a conditional statement (which produces elements that are either `TRUE` or `FALSE`) and outputs one thing IF the element is `TRUE` and outputs something ELSE if the element is `FALSE`. The function's useage looks like this:

`if_else`(conditional statement, output if TRUE, output if FALSE)

For example, to evaluate whether a major is a Science major, we can use the conditional statement:

```
stem_type == "Science"
```

When we are creating the dummy variable `science`, we will give this variable a value of `1` if the STEM category is Science (a `TRUE` element in our logical vector) and a `0` if the STEM category is not Science (a `FALSE` element in our logical vector).

The full `if_else()` syntax to create a `science` dummy-coded variable is this:

```
# Create science dummy variable
stem = stem %>%
  mutate( science = if_else(stem_type == "Science", 1, 0) )

# Examine data
head(stem)
```

```
# A tibble: 6 x 5
                               major income    women   stem_type science
                               <chr>  <dbl>    <dbl>       <chr>   <dbl>
1                Aerospace Engineering     60 13.97928 Engineering       0
2                  Applied Mathematics     45 43.42984 Mathematics       0
3            Architectural Engineering     54 35.04425 Engineering       0
4                         Architecture     40 45.14649 Engineering       0
5           Astronomy And Astrophysics     62 53.57143     Science       1
6 Atmospheric Sciences And Meteorology     35 32.12961     Science       1
```

Since the fifth and sixth majors in the dataset had a `stem_type` value that was `Science`, the dummy code for each of them is 1. The dummy code for all other STEM categories will be 0. Here we will create all four dummy variables. All four can be put in the same `mutate()` layer.

```
# Create all four dummy variables
stem = stem %>%
  mutate(
    science = if_else(stem_type == "Science", 1, 0),
    tech = if_else(stem_type == "Technology", 1, 0),
    engineer = if_else(stem_type == "Engineering", 1, 0),
    math = if_else(stem_type == "Mathematics", 1, 0)
    )

# Examine data
head(stem)
```

```
# A tibble: 6 x 8
                               major income    women   stem_type science
                               <chr>  <dbl>    <dbl>       <chr>   <dbl>
1                Aerospace Engineering     60 13.97928 Engineering       0
2                  Applied Mathematics     45 43.42984 Mathematics       0
3            Architectural Engineering     54 35.04425 Engineering       0
4                         Architecture     40 45.14649 Engineering       0
5           Astronomy And Astrophysics     62 53.57143     Science       1
6 Atmospheric Sciences And Meteorology     35 32.12961     Science       1
# ... with 3 more variables: tech <dbl>, engineer <dbl>, math <dbl>
```

If you do not know the actual names of the categories (or you want to check capitalization, etc.) use the `unique()` function to obtain the unique category names.

```
# Get the categories
unique(stem$stem_type)
```

```
[1] "Engineering" "Mathematics" "Science"     "Technology"
```

Once the dummy variables have been created, fit the regression using all but one of the dummy variables you created. The dummy variable you leave out will correspond to the reference category. For example, in the model fitted below, we include the predictors `tech`, `engineer`, and `math` as prdictors in the model; we did not include the `science` predictor. As such, the STEM category of Science is our reference group.

```
# science is reference group
lm.science = lm(income ~ 1 + tech + engineer + math, data = stem)
summary(lm.science)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.075      1.631  23.346  < 2e-16 ***
tech           3.811      4.042   0.943    0.349
engineer      19.308      2.442   7.907 2.28e-11 ***
math           6.175      5.157   1.197    0.235
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.786 on 72 degrees of freedom
Multiple R-squared:  0.4714,    Adjusted R-squared:  0.4494
F-statistic:  21.4 on 3 and 72 DF,  p-value: 5.144e-10
```

At the model-level, differences in STEM categories seem to explain a statistically significant amount of the variation in reading scores ($F(3, 72) = 21.4$, $p < .0001$). In fact, differences in STEM category explain 47.14% of the variation in median incomes.

At the coefficient-level, the intercept is the average Y value for the reference group. Each partial slope is the difference in average Y values between the reference group and the group represented by the dummy variable. In our example,

- The average income for students in a science major is $38,075.
- Students in a technology major earn $3,811 more a year, on average, than students who earn a science major.
- Students in an engineering major earn $19,308 more a year, on average, than students who earn a science major.
- Students in a mathematics major earn $6,175 more a year, on average, than students who earn a science major.

It is important to note that the partial slope associated with the difference between science and technology majors ($p = .349$) and that between mathematics and science majors ($p = .235$) are not statistically significant. This implies that there is likely no difference in average income between science majors and technology and mathematics majors. The partial slope for the difference between science and engineering majors ($p < .001$), however, indicates that there is a population difference in the average income between these STEM majors.

Omnibus Test vs. Coefficient Tests with Multiple Dummy Variables

When we use multiple dummy variables to represent a single categorical predictor, each β-term represents the mean difference between two groups. For example, in our model,

$$\begin{aligned}\beta_1 &= \mu_{\text{Technology}} - \mu_{\text{Science}} \\ \beta_2 &= \mu_{\text{Engineering}} - \mu_{\text{Science}} \\ \beta_3 &= \mu_{\text{Mathematics}} - \mu_{\text{Science}}\end{aligned}$$

Recall that one manner in which we could write the null hypothesis associated with the model-level test is:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

When we express the null hypothesis at the model-level when we use multiple dummy variables to represent a single categorical predictor, the test includes the mean differences between ALL groups, not just the differences included in the model. In our example, it represents

$$\begin{aligned}\beta_1 &= \mu_{\text{Technology}} - \mu_{\text{Science}} \\ \beta_2 &= \mu_{\text{Engineering}} - \mu_{\text{Science}} \\ \beta_3 &= \mu_{\text{Mathematics}} - \mu_{\text{Science}} \\ \beta_4 &= \mu_{\text{Engineering}} - \mu_{\text{Technology}} \\ \beta_5 &= \mu_{\text{Mathematics}} - \mu_{\text{Technology}} \\ \beta_6 &= \mu_{\text{Engineering}} - \mu_{\text{Mathematics}}\end{aligned}$$

We can express this as

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

or as

$$\begin{aligned}H_0 : &\left(\mu_{\text{Technology}} - \mu_{\text{Science}}\right) = \left(\mu_{\text{Engineering}} - \mu_{\text{Science}}\right) = \left(\mu_{\text{Mathematics}} - \mu_{\text{Science}}\right) = \\ &\left(\mu_{\text{Engineering}} - \mu_{\text{Technology}}\right) = \left(\mu_{\text{Mathematics}} - \mu_{\text{Technology}}\right) = \left(\mu_{\text{Engineering}} - \mu_{\text{Mathematics}}\right) = 0\end{aligned}$$

The test at the model-level is considering all six differences simultaneously. If the model-level test is significant, it is important to examine all potential coefficient-level differences, not just those outputted from the fitted model.

Link to the ANOVA Test

Note that if all the means are equal, then each difference in the previous hypothesis would be 0. So we could also write the model-level null hypothesis as,

$$H_0 : \mu_{\text{Science}} = \mu_{\text{Technology}} = \mu_{\text{Engineering}} = \mu_{\text{Mathematics}}$$

This is the omnibus null hypothesis associated with the one-factor ANOVA. Fitting a regression model with dummy-variables is the same analysis as carrying out an ANOVA. The difference is that the output from the multiple regression gives β-terms associated with mean differences (to the reference group), and ANOVA is concerned more directly with the group mean. But the model-level regression results are identical to those from the ANOVA. Asking whether the model explains variation in the outcome (H_0"$\rho^2 = 0$) is the same as asking whether there are mean differences ($H_0 : \mu_{\text{Science}} = \mu_{\text{Technology}} = \mu_{\text{Engineering}} = \mu_{\text{Mathematics}}$); these are just different ways of writing the model-level null hypothesis!

Further Understanding Income Differences

If you are only interested in if there are differences, you can focus on the model-level (omnibus) results. If, however, you want to go further and further examine the differences between STEM categories, we need to look at the coefficient-level results. Based on the model fitted so far, we have considered only three of the six possible differences.

Comparison	Mean Difference	p
Science − Technology	−$3,811	0.349
Science − Engineering	−$19,308	<0.001
Science − Mathematics	−$6,175	0.235
Technology − Engineering	?	?
Technology − Mathematics	?	?
Engineering − Mathematics	?	?

In order to examine the remaining differences, we need to fit additional regression models that allow for those comparisons. Below, we fit the three other models that use three of the four dummy-coded STEM categories to predict variation in income. (The R output from these models is not shown, but the pertinent results are presented in Table 1.)

```
# tech is reference group
lm.tech = lm(income ~ 1 + science + engineer + math, data = stem)

# math is reference group
lm.math = lm(income ~ 1 + science + tech + engineer, data = stem)

# engineer is reference group
lm.engineer = lm(income ~ 1 + science + tech + math, data = stem)
```

At the model-level, all four models give the same information: Differences in STEM categories explain a statistically significant amount of the variation in incomes, $F(3, 72) = 21.4$, $p < .0001$. In fact, differences in STEM category explain 47.14% of the variation in median incomes.

Using the coefficient-level output from the models, we can fill in the remaining cells of the table. Based on the results, it looks as though there are statistically significant income differences between engineering majors and each of the other types of STEM majors.

Table 1

Mean Income Difference (and p-Value) Between STEM Categories

Comparison	Mean Difference	p
Science − Technology	−$3,811	0.349
Science − Engineering	−$19,308	<0.001
Science − Mathematics	−$6,175	0.235
Technology − Engineering	−$15,497	<0.001
Technology − Mathematics	−$2,364	0.701
Engineering − Mathematics	−$13,133	0.014

Table 2

Regression Results from Fitting Four Different Models to Predict Median Income from a Set of Dummy-Coded STEM Categories

	lm.science	lm.tech	lm.math	lm.engineer
Technology	3.811		−2.364	−15.497***
	(4.042)		(6.133)	(4.121)
Science		−3.811	−6.175	−19.308***
		(4.042)	(5.157)	(2.442)
Engineering	19.308***	15.497***	13.133**	
	(2.442)	(4.121)	(5.219)	
Mathematics	6.175	2.364		−13.133**
	(5.157)	(6.133)		(5.219)
Constant	38.075***	41.886***	44.250***	57.383***
	(1.631)	(3.699)	(4.893)	(1.817)
R^2	0.471	0.471	0.471	0.471
F Statistic (df = 3; 72)	21.402***	21.402***	21.402***	21.402***

Note: *p<0.1; **p<0.05; ***p<0.01

Multiple Comparisons

When we evaluated the p-values for each of these differences, we used an alpha value of 0.05 as the criterion for significance. This is consistent with how we have evaluated other predictors in regression models. This is okay when the effect constitutes a single term or mean difference in the null hypothesis. For a predictor with more than two levels, however, the null hypothesis constitutes more than one mean difference.

For the effect of STEM category, we really have six differences. To be "fair" with other predictors we might include in the model that would constitute a single term/difference, we should really split the 0.05 across the six differences. The easeist manner to make this "fair" is to divide the 0.05 evenly across the six differences.

$$\frac{0.05}{6} = 0.0083$$

Then, rather than rejecting the null hypothesis when the p-value is below 0.05, we will only reject if the comparison has a p-value below 0.0083. Looking back at the table we created earlier, the income differnces between mathematics and engineering majors is no longer significant ($p = .014 \not< .0083$).

Adjusting the p-Value (not alpha)

In practice, people are psychologically accustomed to comparing the p-value to 0.05, so changing the alpha-value to .0083 can be a problem. Another way to achieve the same adjustment, but still allow people to compare to 0.05, is to change the p-value rather than change the alpha value. To do this, we multiply each p-value by 6 (rather than dividing the alpha value by 6).

```
p_values = c(
  0.349,           # Science vs. Technology
  0.0000000000228, # Science vs. Engineering
  0.235,           # Science vs. Mathematics
  0.000342,        # Technology vs. Engineering
  0.701,           # Technology vs. Mathematics
  0.014            # Engineering vs. Mathematics
)

p_values * 6
```

```
[1] 2.094e+00 1.368e-10 1.410e+00 2.052e-03 4.206e+00 8.400e-02
```

Table 3

Unadjusted and Bonferroni Adjusted p-Values for Income Comparisons Between STEM Categories

Difference	Unadjusted p	Adjusted p
Science − Technology	0.349	2.094
Science − Engineering	0.0000000000228	<0.001
Science − Mathematics	0.235	1.410
Technology − Engineering	0.000342	0.002
Technology − Mathematics	0.701	4.206
Engineering − Mathematics	0.014	0.084

When reporting the adjusted p-values, be careful. Remember that p-values are always between 0 and 1. Anything above 1 needs to be reported as 1! This method of evenly splitting the alpha value or adjusting the p-value evenly is called the *Bonferroni adjustment*.

We can also use the `p.adjust()` function to compute the Bonferroni adjusted p-values. To use this, create a vector of the unadjusted p-values and then include this vector in the `p.adjust()` function along with the argument `method = "bonferroni"`.

```
# Bonferroni adjustment to the p-values
p.adjust(p_values, method = "bonferroni")
```

```
[1] 1.000e+00 1.368e-10 1.000e+00 2.052e-03 1.000e+00 8.400e-02
```

Table 4

Unadjusted and Bonferroni Adjusted p-Values for Income Comparisons Between STEM Categories

Difference	Unadjusted p	Adjusted p
Science − Technology	0.349	1.000
Science − Engineering	<0.001	<0.001
Science − Mathematics	0.235	1.000
Technology − Engineering	<0.001	0.002
Technology − Mathematics	0.701	1.000
Engineering − Mathematics	0.014	0.084

Other p-Value Adjustment Methods

There is nothing that requires you to evenly adjust the p-value across the six comparisons. For example, some adjustment methods use different multipliers depending on the size of the initial unadjusted p-value. One of those methods is the *Benjamini–Hochberg adjustment*. This adjustment procedure ranks the unadjusted p-values from smallest to largest and then adjusts by the following computation[1]:

$$p_{\text{adjusted}} = \frac{k \times p_{\text{unadjusted}}}{\text{Rank}}$$

In this adjustment, the numerator is equivalent to making the Bonferroni adjustment. The size of the Bonferroni adjustment is then scaled back depending on the initial rank of the unadjusted p-value. The smallest initial p-value gets the complete Bonferroni adjustment, while the largest Bonferroni adjustment is scaled back the most. We can use `method="BH"` in the `p.adjust()` function to obtain the Benjamini–Hochberg adjusted p-values directly.

```
# Benjamini-Hochberg adjustment to the p-values
p.adjust(p_values, method = "BH")
```

```
[1] 4.188e-01 1.368e-10 3.525e-01 1.026e-03 7.010e-01 2.800e-02
```

Table 5

Unadjusted and Benjamini–Hochberg Adjusted p-Values for Income Comparisons Between STEM Categories

Difference	Unadjusted p	Adjusted p
Science − Technology	0.349	0.419
Science − Engineering	0.0000000000228	<0.001
Science − Mathematics	0.235	0.353
Technology − Engineering	0.000342	0.001
Technology − Mathematics	0.701	0.701
Engineering − Mathematics	0.014	0.028

[1]The actual adjusted p-value given is the minimum of this value and the adjusted p-value for the next higher raw p-value.

Using the Benjamini–Hochberg adjusted p-values, we find statistically significant income differences between (1) science and engineering majors, (2) technology and engineering majors, and (3) engineering and mathematics majors.

Which Adjustment Method?

There are many, many different adjustment methods you can choose. The `p.adjust()` function, for example, includes six adjustment options (the Holm method, the Hochberg method, the Hommel method, the Bonferroni method, the Bnjamani–Hochberg method, and the Benjamini–Yekutieli method). In addition, the **multcomp** package includes several other adjustment methods.

You should decide which adjustment method you will use before you do the analysis. In the social sciences, the Bonferroni method has been historically the most popular method (probably because it was easy to implement before computing). That being said, I would encourage you to use the Benjamini–Hochberg adjustment method. It is from a family of adjustment methods that a growing pool of research evidence points toward as the "best" solution to the problem of multiple comparisons (Williams, Jones, & Tukey, 1999). Because of its usefulness, the Institute of Education Sciences has recommended this procedure for use in its What Works Clearinghouse Handbook of Standards.

Does STEM Type Mediate the Relationship Between Proportion of Women and Income?

According to Wikipedia,

> [A] mediation model is one that seeks to identify and explain the mechanism or process that underlies an observed relationship between an independent variable and a dependent variable via the inclusion of a third hypothetical variable, known as a mediator variable... Rather than a direct causal relationship between the independent variable and the dependent variable, a mediation model proposes that the independent variable influences the (non-observable) mediator variable, which in turn influences the dependent variable.

In our example, we hypothesize that is is not the influx of women into a major that causes lower median incomes, but rather that women are attracted to the "S"-majors and it is the type of STEM major that is causing the lower incomes. IN this sense, we would argue that STEM type MEDIATES the relationship between the proportion of women graduating with that major and the income level.

To test this hypothesis, we will fit a multiple regression model that includes both the proportion of women graduating and the set of dummy-coded STEM type predictors; a multiple regression model. Then, we can see if the partial/controlled relationship associated with the `women` predictor has changed from the simple regression model.

```
lm.mediator = lm(income ~ 1 + women + tech + engineer + math, data = stem)
summary(lm.mediator)
```

```
Call:
lm(formula = income ~ 1 + women + tech + engineer + math, data = stem)

Residuals:
    Min      1Q  Median      3Q     Max
-19.562  -5.137  -1.008   2.247  51.020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.64140    5.66986   8.226 6.36e-12 ***
women       -0.13497    0.08563  -1.576  0.11945
tech        -1.20483    5.11237  -0.236  0.81437
engineer    13.96556    4.16296   3.355  0.00128 **
math         2.96629    5.49608   0.540  0.59109
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.686 on 71 degrees of freedom
Multiple R-squared:  0.4893,    Adjusted R-squared:  0.4605
F-statistic:    17 on 4 and 71 DF,  p-value: 8.042e-10
```

At the model-level, we find that differences in the proportion of women who graduate from the major and STEM category explain 48.93% of the variation in median incomes. This is a statistically significant amount of variation, $F(4, 71) = 17.0$, $p < .0001$.

At the coefficient-level, we are most interested in the β-coefficient associated with the `women` predictor. This coefficient, $\hat{\beta} = -0.13$, indicates that after controling for type of STEM major, majors that graduate more women have lower median incomes, on average. Each one-percent difference in the percentage of female graduates is associated with a $134 decrease in median income, on average.

The more interesting piece is that this effect, after controlling for differences in the type of STEM major, is NOT statistically significant, $t(71) = -1.58$, $p = .119$. Recall when we examined this effect from the simple regression model, it was statistically significant. After controlling for type of STEM major, the effect has gone away. This seems to support our mediation hypothesis that the type of STEM major mediates the relationship between percentage of female graduates and median income.

Mediation is really a hypothesis about the underlying causal mechanism. Understanding the nature of cause is a substantive, not statistical, issue. The way we analyze mediation is by comparing uncontrolled effects (simple regression) to controlled effects (multiple regression). As Kenny (2016) writes, "if the presumed causal model is not correct, the results from the mediational analysis are likely of little value".

ANCOVA: Controlled Group Differences

Sometimes the goal of the analysis is the group differences. For example, our focal research might be on the income differences between the categories of STEM majors. In this case, we would likely be interested in evaluating whether the differences we saw in the uncontrolled model persist after controlling for differences in one or more covariates. In psychology, an analysis that focuses on controlled group differences is referred to as an *Analysis of Covariance* or ANCOVA.

I will use the STEM data to illustrate ANCOVA. We will examine whether income difference across STEM type persist after controlling for the percentage of women in the major. In this analysis, the focus is on the group differences, NOT on the `women` effect. We can again fill out a table of mean differences, unadjusted and adjusted p-values. But for the controlled group differences, we will use the results from our multiple regression analysis. (The syntax for fitting the models is shown below, but the summary results are not printed.)

```
# Fit ANCOVA models
lm.science.control  = lm(income ~ 1 + women +           tech + engineer + math, data = stem)
lm.tech.control     = lm(income ~ 1 + women + science +        engineer + math, data = stem)
lm.engineer.control = lm(income ~ 1 + women + science + tech +             math, data = stem)
lm.math.control     = lm(income ~ 1 + women + science + tech + engineer,        data = stem)

# Adjust p-values
p_values = c(0.814, 0.001, 0.591, 0.0004, 0.501, 0.043)
p.adjust(p_values, method = "BH")
```

```
[1] 0.8140 0.0030 0.7092 0.0024 0.7092 0.0860
```

Table 6

Adjusted Mean Differences, Unadjusted and Benjamini–Hochberg Adjusted p-Values for Income Comparisons Between STEM Categories Controlling for Differences in the Percentage of Female Graduates

Difference	Adj. Mean Difference	Unadjusted p	Adjusted p
Science − Technology	1,205	0.814	0.814
Science − Engineering	−13,966	0.001	0.003
Science − Mathematics	−2.966	0.591	0.709
Technology − Engineering	−15,170	0.0004	0.002
Technology − Mathematics	−4,171	0.501	0.709
Engineering − Mathematics	10,999	0.043	0.086

Using the Benjamini–Hochberg adjusted p-values, after controlling for differences in the percentage of female graduate, we find statistically significant income differences between (1) science and engineering majors, and (2) technology and engineering majors. The income differences we initially observed between engineering and mathematics majors has disappeared.

In the language of ANCOVA, the controlled mean differences are referred to as *Adjusted Mean Differences*. So, for example, the adjusted mean difference in incomes between science and technology majors is $1,205 (controlling for differences in the percentage of female graduates). When the mean difference is from a model that has no covariates, it is referred to as an *Unadjusted Mean Difference*. It can be useful to present both the unadjusted and adjusted mean differences in a table.

Table 7

Unadjusted and Adjusted Mean Differences for Income Comparisons Between STEM Categories. Adjusted Mean Differences are Controlling for Differences in the Percentage of Female Graduates

	Mean Difference	
Difference	Unadjusted	Adjusted
Science − Technology	−3811	1205
Science − Engineering	−19 308***	−13 966**
Science − Mathematics	−6175	−2966
Technology − Engineering	−15 497**	−15 170**
Technology − Mathematics	−2364	−4171
Engineering − Mathematics	−13 133*	10 999

Note. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

The results show how the income difference between STEM categories changes when we control for differences in other covariates, in this case, the percentage of female graduates.

Technical Reasons to Adjust for Multiple Comparisons

In the earlier sections, we presented the reason for adjusting the p-values for the STEM category comparisons as one of "fairness" with the other predictors in the model. This is true, but there are also technical reasons to make these adjustments. The main technical reason is related to the *Type I error rate*. Remember that a Type I error occurs when you falsely reject a true null hypothesis. In other words, we would say there is an income difference between STEM categories when there really isn't a difference.

When we use an alpha value of 0.05, we are saying we are willing to make a Type I error in 5% of the samples that could be randomly selected (we have no idea whether our sample is one of the 5% where we will make an error, or one of the 95% where we won't). For effects that only have one row in the model, there is only one test in which we can make a Type I error ($H_0 : \beta_j = 0$), so we are okay evaluating each at the alpha of 0.05.

WHen we have more than two levels of a categorical predictor, there are multiple differences that constitute the effect of that predictor. To test whether there is an effect of that predictor, we evaluate multiple hypotheis tests. For our STEM data, to test whether there is an effect of STEM category on income, we evaluate six hypothesis tests:

$$
\begin{aligned}
H_0 :& \mu_{\text{Technology}} - \mu_{\text{Science}} = 0 \\
H_0 :& \mu_{\text{Engineering}} - \mu_{\text{Science}} = 0 \\
H_0 :& \mu_{\text{Mathematics}} - \mu_{\text{Science}} = 0 \\
H_0 :& \mu_{\text{Engineering}} - \mu_{\text{Technology}} = 0 \\
H_0 :& \mu_{\text{Mathematics}} - \mu_{\text{Technology}} = 0 \\
H_0 :& \mu_{\text{Engineering}} - \mu_{\text{Mathematics}} = 0
\end{aligned}
$$

Because of this, there are many ways to make a Type I error. For example, we could make a Type I error in any one of the six tests, or in two of the six tests, or in three of the six tests, etc. Therefore, the probability of making at least one Type I error is no longer 0.05, it is

$$1 - (1 - \alpha)^k$$

where α is the alpha level for each test, and k is the number of tests (comparisons) for the effect.

In our example this is

$$P(\text{type I error}) = 1 - (1 - 0.05)^6 = 0.735$$

The probability that we will make at least one Type I error in the six tests is 73.5%!!! This probability is called the family-wise Type I error rate. In the social sciences, the family-wise error rate needs to be 0.05. What should α be if we want the family-wise error rate to be 0.05? Essentially we would need to solve this equation:

$$0.05 = 1 - (1 - \alpha)^6$$

Carlo Emilio Bonferroni solved this algebra problem for any value of k and found that the value for alpha that $\frac{\text{family-wise error rate}}{k}$ gives an upper-bound for the solution. Olive Jean Dunn then used Bonferroni's solution in practice. This is why dividing by the number of comparisons is referred to as the Bonferroni or the Dunn–Bonferroni method.

False Discovery Rate

The Benjamini–Hochberg procedure is an ensemble method based on *false discovery rate* (FDR). FDR is a relatively new approach to the multiple comparisons problem. Instead of making adjustments to control the probability of making at least one Type I error, FDR controls the *expected proportion of discoveries* (rejected null hypotheses) when the null hypothesis is true; in other words, it controls the expected proportion of Type I error. You can find out more from Wikipedia.

The FDR concept was formally described in a 1995 paper by Yoav Benjamini and Yosi Hochberg, and resulted in their proposal of the Benjamini–Hochbergm method (Benjamini & Hochberg, 1995). They argued that using FDR produces a less conservative and arguably more appropriate approach for identifying statistically significant comparisons.

In practice, using FDR rather than family-wise adjustment of error makes these methods less prone to over-adjustment of the p-values. However, the increased statistical power that comes with using the FDR methods is not without cost. They also have increased probabilities of Type I errors relative to the family-wise adjustment methods.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*(1), 289—300.

Kenny, D. A. (2016). Mediation. Personal website. Retrieved from http://davidakenny.net/cm/mediate.htm

Williams, V., Jones, L., & Tukey, J. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, *24*(1), 42–69.