More Categorical Predictors

2018-07-30

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 college majors. We will use this to explore several speculative hypotheses about the choice of major, including:

- Women are more likely to choose a college major that earns less money.
- STEM majors make more money than non-STEM majors.

The dataset, *stem.csv*, includes data on 172 college majors collected from the *American Community Survey 2010–2012 Public Use Microdata Series* on graduates who were under the age of 29. Variabes in the dataset are:

- `major`: Name of STEM major
- `income`: Median income (in thousands of dollars) for a full-time, year-round worker
- `num_grad`: Number of total graduates
- `women`: Percentage of total graduates who are women
- `unemployment`: Percentage of unemployed graduates
- `major_cat`: Category of major from Carnevale, Strohl, & Melton (2011)
- `stem_major`: Is the major a STEM (Science-Technology-Engineering-Mathematics) major? (0 = No; 1 = Yes)
- `s_tem`: Trichotomization of major categories (Science; Tech-Engineer-Math; Non-STEM)

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

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

```
# A tibble: 6 x 8
  major    income num_grad women unemployment major_cat  stem_major s_tem
  <chr>     <dbl>    <int> <dbl>        <dbl> <chr>           <int> <chr>
1 ACCOUNT~     45   198633  52.4         6.97 Business            0 Non-S~
2 ACTUARI~     62     3777  44.1         9.57 Business            0 Non-S~
3 ADVERTI~     35    53162  75.8         6.80 Communica~          0 Non-S~
4 AEROSPA~     60    15058  14.0         6.52 Engineeri~          1 Tech-~
5 AGRICUL~     40     2439  28.3         7.72 Agricultu~          0 Non-S~
6 AGRICUL~     40    14240  32.2         5.00 Agricultu~          0 Non-S~
```

Examine and Describe the Marginal Distribution of the Median Incomes

To begin the analysis, we will explore the outcome variable, median income.

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

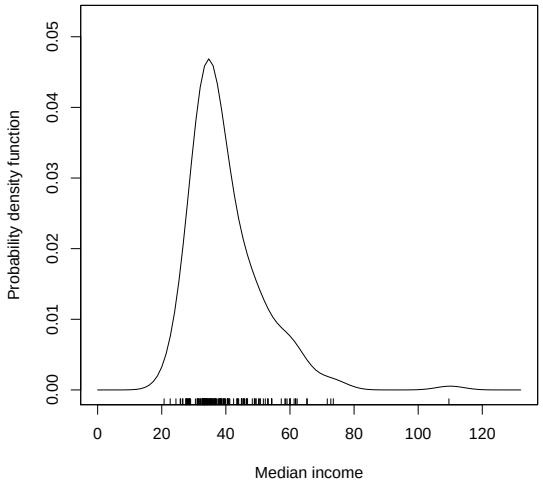

Figure 1: Density plot of the median incomes for $n = 76$ STEM majors.

```
# 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  40.1  11.5    22   110
```

The median incomes for the 172 college majors are right-skewed and range from $22,000 (Library Science) to $110,000 (Petroleum Engineers). The mean income is $40,000. The standard deviation of roughly $11,500 suggests that most graduates earn between $17,000 and $63,000.

To put this in perspective, around the same time period, the Bureau of Labor Statistics estimated that the median income for a person with only a high school education was $25,500.

H1: Women are more likely to choose a college major that earns less money.

To explore this hypothesis, we can examine a scatterplot of the relationship between the percentage of 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)")
```

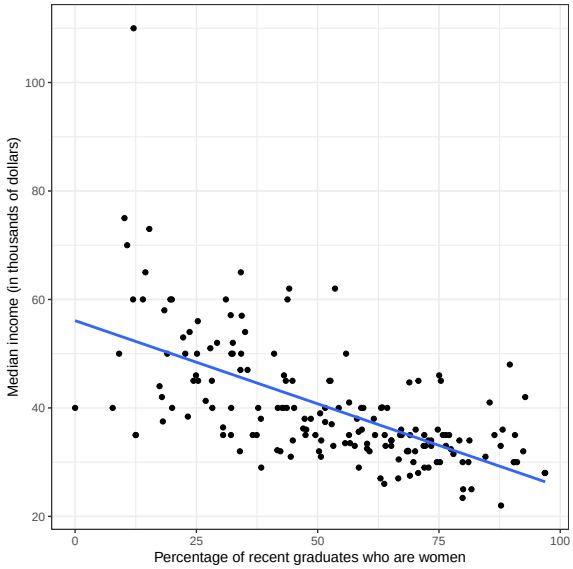

Figure 2: Scatterplot showing the relationship between the percentage of recent graduates in each STEM major that are women, and the median income for those majors. The OLS regression lines is also displayed on the plot.

We also compute the correlation matrix for these two variables.

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

```
  rowname income women
1  income        -.619
2   women  -.619
```

The correlation coefficient and scatterplot suggest a negative relationship between these variables. This implies that majors that have a higher percentage of female graduates tend to be the same majors that have lower median incomes ($r = -.619$). 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

We can also fit a model that regresses median income on percentage of females.

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

# Model-level info
glance(lm.1)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1  0.382777     0.3791463 9.030918  105.4272 1.507304e-19  2 -621.5641
       AIC      BIC deviance df.residual
1 1249.128 1258.571 13864.77         170
```

Differences in the pecentage of females in the major explains 38.2% of the variation in median incomes. This explain variation is statistically diffeent than 0, $F(2, 170) = 105.43$, $p < .001$.

```
# Coefficient-level info
tidy(lm.1)
```

```
         term   estimate std.error statistic      p.value
1 (Intercept) 56.0933051 1.7051148  32.89708 1.266844e-75
2       women -0.3066994 0.0298701 -10.26777 1.507304e-19
```

The fitted regression model is

$$\hat{\text{Median Income}} = 56.09 - 0.31(\text{Percentage of Women})$$

- The fitted intercept ($\hat{\beta}_0 = 56.09$) indicates that the median income for college majors that are 100% male is roughly $56,000, on average.
- The fitted slope ($\hat{\beta}_1 = -0.31$) indicates that, on average, the difference in median incomes for college majors that have one-percent more female graduates is $310.

The p-value associated with the slope suggests that this difference in income is statistically significant ($p < .001$).

H2: STEM majors make more money than non-STEM majors.

To explore the second speculative hypothesis, we will examine a scatterplot and correlation matrix of the median incomes versus whether or not the major is a STEM major. For better plotting, we will coerce `stem` into a factor in the `ggplot()` function.

```
# Plot the median incomes by STEM major
ggplot(data = stem, aes(x = factor(stem_major), y = income, fill = factor(stem_major))) +
  geom_point(shape = 21, color = "black", size = 4) +
    theme_bw() +
    scale_x_discrete(name = "", labels = c("Non-STEM major", "STEM major")) +
    ylab("Median income (in thousands of dollars)") +
  scale_fill_viridis_d() +
  guides(fill = FALSE)
```

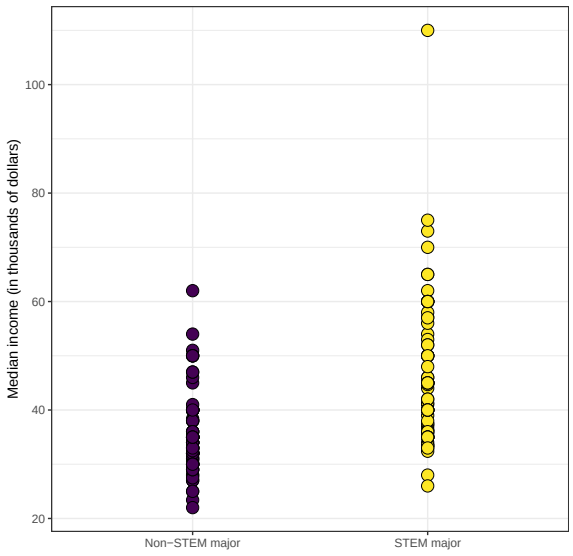

Figure 3: Scatterplot showing median income versus whether or not the major is a STEM major.

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

```
# A tibble: 2 x 4
  stem_major     M    SD     N
       <int> <dbl> <dbl> <int>
1          0  35.3  6.84    96
2          1  46.1 13.2     76
```

We also compute the correlation matrix for these two variables.

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

```
     rowname income stem_major
1     income              .470
2 stem_major   .470
```

The data suggests that there are potential income differences between STEM and non-STEM majors. On average, STEM majors earn about $11,000 more annually than non-STEM majors. However, there is a great deal of variation in median incomes for both groups.

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 package
library(ggridges)

# Ridge plot
# - coerce stem_major into a factor
# - The variable mapped to x= has to be continuous, use coord_flip() to
#   emulate the positioning on the scatterplot
ggplot(data = stem, aes(x = income, y = factor(stem_major))) +
  geom_density_ridges() +
  theme_bw() +
  xlab("Median income") +
  ylab("Dummy coded STEM major") +
  coord_flip()
```

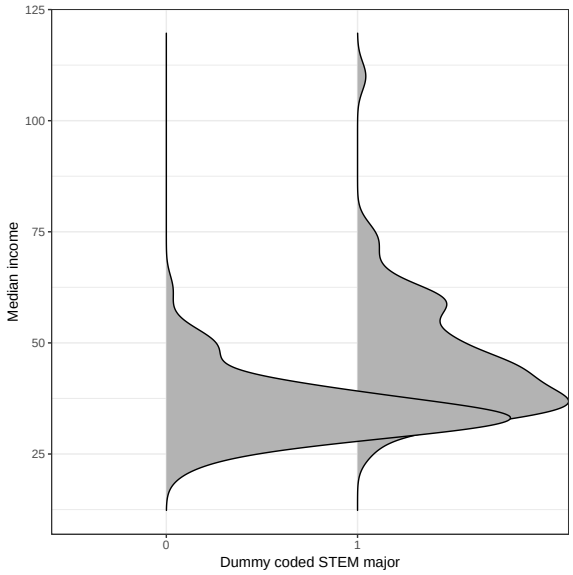

Figure 4: Ridge plot showing the distribution of median income for STEM and non-STEM majors.

This plot suggests the same pattern in medain incomes as the scatterplot: STEM majors earn a higher salary on average than their non-STEM peers.

Fit the Regression Model

To examine whether the observed difference in income is due ti chance, we fit a model regressing median income on the dummy variable `stem_major`.

```
lm.2 = lm(income ~ 1 + stem_major, data = stem)

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

```
  r.squared adj.r.squared    sigma statistic     p.value df    logLik
1 0.2212663     0.2166855 10.14391  48.30311 7.46804e-11  2 -641.5538
       AIC     BIC deviance df.residual
1 1289.108 1298.55 17492.81         170
```

At the model-level, differences in major categorization seem to explain 22.1% of the variation in median incomes; a statistically significant amount of the variation ($F(2, 170) = 48.30$, $p < .001$).

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

```
         term estimate std.error statistic      p.value
1 (Intercept) 35.29375  1.035308 34.090085 6.560582e-78
2  stem_major 10.82467  1.557497  6.950044 7.468040e-11
```

The fitted regression equation is

$$\widehat{\text{Median Income}} = 35.294 + 10.825(\text{STEM Major})$$

- The fitted intercept ($\hat{\beta}_0 = 35.294$) indicates that the median income for non-STEM majors is roughly \$35,000, on average.
- The fitted slope ($\hat{\beta}_1 = 10.825$) indicates that, on average, STEM majors have a median income that is \$10,800 higher than their non-STEM peers, on average.

The p-value associated with the slope suggests that this income difference is statistically significant ($p < .001$); likely not due to chance.

H3: "TEM" majors are more coveted than "S" majors.

More recently, it has been suggested that not all STEM majors are created equal. In the fivethirtyeight.com article The Economic Guide To Picking A College Major, Ben Casselman wrote:

> Politicians love to tout the importance of science, technology, engineering and math majors. But when it comes to earnings, the "S" majors don't really belong with the "TEM" ones.

To explore this hypothesis, we can plot the median incomes versus the `s_tem` categorization.

```
# Ridge plot
ggplot(data = stem, aes(x = income, y = s_tem)) +
  geom_density_ridges() +
  theme_bw() +
  xlab("Median income") +
  ylab("Major categorization") +
  coord_flip()
```

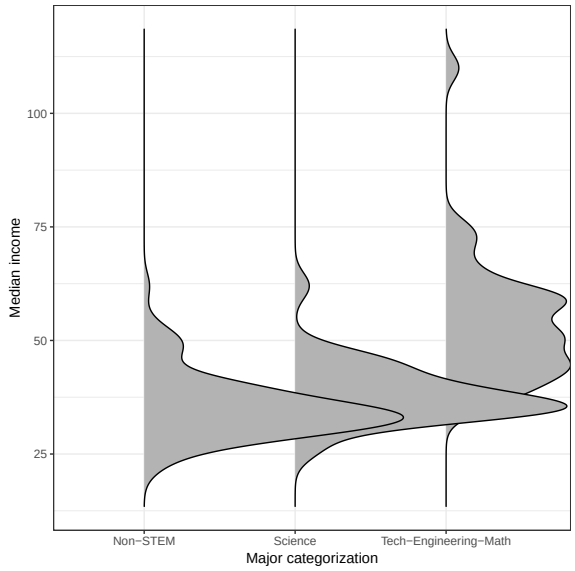

Figure 5: Ridge plot showing the distribution of median income for "S", "TEM", and non-STEM majors.

The plots for the Science ("S")-majors and non-STEM majors look relatively similar. The distribution of incomes associated with the Technolgy, Engineering, and Mathematics ("TEM")-majors looks to have a higher median income than the other two groups.

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

```
# A tibble: 3 x 4
  s_tem                    M    SD     N
  <chr>                <dbl> <dbl> <int>
1 Non-STEM              35.3  6.84    96
2 Science               38.1  6.43    36
3 Tech-Engineering-Math  53.4 13.6     40
```

The mean income for "S"-majors is lower than for "TEM"-majors, and comparable to that for non-STEM majors. The "TEM" majors earn roughly $15,000 more than their peers.

Fitting a Regression Model

To examine whether the differences we observed are due to chance, we need to fit a model regressing median income on the `s_tem` categorization. Before fitting this model, we need to create a dummy variable for EACH category of the `s_tem` variable. For our analysis, we will need to create three dummy variables: `science`, `tech_eng_math`, and `non_stem`. 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:

```
s_tem == "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 %>%
  mutate(
    science = if_else(s_tem == "Science", 1, 0)
    )
```

```
# A tibble: 172 x 9
   major income num_grad women unemployment major_cat stem_major s_tem
   <chr>  <dbl>    <int> <dbl>        <dbl> <chr>           <int> <chr>
 1 ACCO~     45   198633  52.4         6.97 Business            0 Non-~
 2 ACTU~     62     3777  44.1         9.57 Business            0 Non-~
 3 ADVE~     35    53162  75.8         6.80 Communic~           0 Non-~
 4 AERO~     60    15058  14.0         6.52 Engineer~           1 Tech~
 5 AGRI~     40     2439  28.3         7.72 Agricult~           0 Non-~
 6 AGRI~     40    14240  32.2         5.00 Agricult~           0 Non-~
 7 ANIM~     30    21573  75.2         5.09 Agricult~           0 Non-~
 8 ANTH~     28    38844  70.7        10.3  Humaniti~           0 Non-~
 9 APPL~     45     4939  43.4         9.08 Computer~           1 Tech~
10 ARCH~     54     2825  35.0         6.19 Engineer~           1 Tech~
# ... with 162 more rows, and 1 more variable: science <dbl>
```

All majors with a `s_tem` value that was `Science`, will have the dummy code 1 in the new `science` variable. The dummy code for all other major categories will be 0. Here we will create all three dummy variables. All three can be put in the same `mutate()` layer. We re-assign this into an object called `stem`

```
# Create all three dummy variables
stem = stem %>%
  mutate(
    science       = if_else(s_tem == "Science",               1, 0),
    tech_eng_math = if_else(s_tem == "Tech-Engineering-Math", 1, 0),
    non_stem      = if_else(s_tem == "Non-STEM",              1, 0)
    )

# Examine data
head(stem)
```

```
# A tibble: 6 x 11
  major income num_grad women unemployment major_cat stem_major s_tem
```

```
  <chr>  <dbl>    <int> <dbl>        <dbl> <chr>          <int> <chr>
1 ACCO~     45   198633  52.4         6.97 Business           0 Non-~
2 ACTU~     62     3777  44.1         9.57 Business           0 Non-~
3 ADVE~     35    53162  75.8         6.80 Communic~          0 Non-~
4 AERO~     60    15058  14.0         6.52 Engineer~          1 Tech~
5 AGRI~     40     2439  28.3         7.72 Agricult~          0 Non-~
6 AGRI~     40    14240  32.2         5.00 Agricult~          0 Non-~
# ... with 3 more variables: science <dbl>, tech_eng_math <dbl>,
#   non_stem <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$s_tem)
```

```
[1] "Non-STEM"              "Tech-Engineering-Math" "Science"
```

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 `science`, and `tech_eng_math` as prdictors in the model; we did not include the `non_stem` predictor. As such, non-STEM majors is our reference group.

```
# non-STEM is reference group
lm.3 = lm(income ~ 1 + science + tech_eng_math, data = stem)

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

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1 0.4182667     0.4113822 8.793341  60.75556 1.315752e-20  3 -616.4713
       AIC      BIC deviance df.residual
1 1240.943 1253.533 13067.56         169
```

At the model-level, differences in major categorization seem to explain a statistically significant amount of the variation in median incomes ($F(3, 169) = 60.75$, $p < .001$). In this model, differences in major categorization explain 41.8% of the variation in median incomes.

Note the explained variation in this model is much higher than the explained variation when we only used two categorizations of major ($R^2 = 0.22$). Increasing variation in the predictor can often lead to stronger explanatory models. This means that if you have a continuous predictor you may not want to cut it up into categories.

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

```
          term estimate std.error statistic      p.value
1   (Intercept) 35.29375 0.8974666 39.325975 5.745352e-87
2       science  2.78125 1.7185178  1.618401 1.074411e-01
3 tech_eng_math 18.06375 1.6548467 10.915664 2.477524e-21
```

The fitted regression equation is

$$\widehat{\text{Median Income}} = 35.294 + 2.781(\text{Science Major}) + 18.063(\text{TEM Major})$$

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 non-STEM majors is \$35,294.
- Students who earned a Science major earn \$2,781 more annually, on average, than their non-STEM peers.
- Students who earned a technology, engineering, or mathematics major earn \$18,063 more annually, on average, than their non-STEM peers.

It is important to note that the partial slope associated with the difference between non-STEM and science majors ($p = .107$) is not statistically significant. This implies that there is likely no difference in average income between science majors and non-STEM majors. The partial slope for the difference between technology, engineering, and mathematics majors and non-STEM majors ($p < .001$), however, indicates that there is a statistically signiicant difference in the average income between these types of 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 fitted equation we see the results of testing the following two hypotheses:

$$\beta_1 = \mu_{\text{Science}} - \mu_{\text{non-STEM}}$$
$$\beta_2 = \mu_{\text{Tech-Eng-Math}} - \mu_{\text{non-STEM}}$$

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

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

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 sets of two groups, not just the differences included in the fitted equation. In our example, it represents

$$\beta_1 = \mu_{\text{Science}} - \mu_{\text{non-STEM}}$$
$$\beta_2 = \mu_{\text{Tech-Eng-Math}} - \mu_{\text{non-STEM}}$$
$$\beta_3 = \mu_{\text{Tech-Eng-Math}} - \mu_{\text{Science}}$$

We can express this as

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

or as

$$H_0 : \left(\mu_{\text{Science}} - \mu_{\text{non-STEM}}\right) = \left(\mu_{\text{Tech-Eng-Math}} - \mu_{\text{non-STEM}}\right) = \left(\mu_{\text{Tech-Eng-Math}} - \mu_{\text{Science}}\right) = 0$$

The test at the model-level is considering all three pairwise differences simultaneously. If the model-level test is significant, any one (or more than one) of the differences may not be zero. Because of this, it is important to examine ALL potential coefficient-level differences, not just those outputted from the initial 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,

$$H_0 : \mu_{\text{non-STEM}} = \mu_{\text{Science}} = \mu_{\text{Tech-Eng-Math}}$$

This is the omnibus null hypothesis associated with the one-factor analysis of variance (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 means. 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 ($H_0 : \mu_{\text{non-STEM}} = \mu_{\text{Science}} = \mu_{\text{Tech-Eng-Math}}$); 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 examine the pairwise differences between major categories, we need to look at the coefficient-level results. Based on the fitted equation from above, so far we have considered only two of the three possible pairwise differences.

Table 1: Pairwise Comparisons between Three College Major Categorizations

Comparison	Mean Difference	p
Science − non-STEM	\$2,781	0.107
Tech/Eng/Math − non-STEM	\$18,064	<0.001
Tech/Eng/Math − Science	?	?

In order to examine the remaining pairwise difference, we need to fit an additional regression model that allows us to evaluate this comparison. Below, we fit a second model (using science majors as the reference group) to predict variation in income.

```
# Science majors is reference group
lm.science = lm(income ~ 1 + non_stem + tech_eng_math, data = stem)

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

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1 0.4182667     0.4113822 8.793341  60.75556 1.315752e-20  3 -616.4713
       AIC      BIC deviance df.residual
1 1240.943 1253.533 13067.56         169
```

Note that the model-level output for this fitted model is exactly the same as that for the model in which non-STEM majors was the reference group. This is because we are fitting the exact same omnibus model (to examine whether the three categorizations explain variation in income).

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

```
           term estimate std.error statistic      p.value
1   (Intercept) 38.07500  1.465557 25.979886 6.572226e-61
2      non_stem -2.78125  1.718518 -1.618401 1.074411e-01
3 tech_eng_math 15.28250  2.020131  7.565104 2.373643e-12
```

The fitted regression equation, which is different than the previous fitted equation, is

$$\widehat{\text{Median Income}} = 38.070 - 2.781(\text{non-STEM Major}) + 15.283(\text{TEM Major})$$

- The average income for Science majors is \$38,070.
- Students who earned a non-STEM major earn \$2,781 less annually, on average, than Science majors ($p = .107$).
- Students who earned a technology, engineering, or mathematics major earn \$15,283 more annually, on average, than Science majors ($p < .001$).

For completeness, we also show the syntax to fit the model where Technology, Engineering and Mathematics majors is the reference group. Note that fitting this model does not give us any new information about the overall explained variation or the pairwise comparisons. (It is redundant information to that we have already obtained from fitting the other two models.)

```
# Technology, engineering, and mathematics majors is reference group
lm.tech = lm(income ~ 1 + non_stem + science, data = stem)

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

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1 0.4182667     0.4113822 8.793341  60.75556 1.315752e-20  3 -616.4713
       AIC      BIC deviance df.residual
1 1240.943 1253.533 13067.56         169
```

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

```
         term  estimate std.error  statistic      p.value
1 (Intercept)  53.35750  1.390349  38.377046 2.358136e-85
2    non_stem -18.06375  1.654847 -10.915664 2.477524e-21
3     science -15.28250  2.020131  -7.565104 2.373643e-12
```

At the model-level, all three models give the same information.

Table 2: Regression Results from Fitting Three Different Models to Predict Median Income from a Set of Dummy-Coded Major Categories

	non-STEM	Science	Tech/Eng./Math
(Intercept)	35.29***	38.07***	53.36***
	(0.90)	(1.47)	(1.39)
science	2.78		−15.28***
	(1.72)		(2.02)
tech_eng_math	18.06***	15.28***	
	(1.65)	(2.02)	
non_stem		−2.78	−18.06***
		(1.72)	(1.65)
R^2	0.42	0.42	0.42
F statistic	60.76	60.76	60.76
RMSE	8.79	8.79	8.79

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Using the coefficient-level output from the fitted equations, we can fill in the remaining cells of the table. Based on the results, it looks as though technology, engineering, and mathematics majors have statistically significantly higher incomes, on average, than science majors and non-STEM majors. There appears to be no significant income differences between science majoars and their non-STEM peers.

Table 3: Pairwise Comparisons between Three College Major Categorizations

Comparison	Mean Difference	p
Science − non-STEM	\$2,781	0.107
Tech/Eng/Math − non-STEM	\$18,064	<0.001
Tech/Eng/Math − Science	\$15,283	<0.001

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 statistical significance. This is consistent with how we have evaluated other predictors in regression models. This is okay when the regression 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 major categoryization, we really have three mean 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 three differences. The easeist manner to make this "fair" is to divide the 0.05 evenly across the three differences.

$$\alpha_{\text{Comparison}} = \frac{0.05}{3} = 0.017$$

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 0.017. Looking back at the table we created earlier, none of our decisions about whether differences were due to chance changed. The income differences between

Tech/Eng./Mathematics and Science majors (p = .00000000002373643 < .017) and non-STEM majors (p = .000000000000000000002477524 < .017) are both still significantly different. The income differences between Science nd non-STEM majors (p = .1074411 > .017) is still not statistically reliable.

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 0.017 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 the number of mean differences that constitute the regression effect, rather than dividing the alpha value by this value. In our example, we would multiply each p-value by 3.

```
p_values = c(
  0.00000000002373643,          # Tech/Eng./Mathematics vs. Science
  0.000000000000000000002477524, # Tech/Eng./Mathematics vs. non-STEM
  0.1074411                     # Science vs. non-STEM
)

# Adjust p-values
p_values * 3
```

```
[1] 7.120929e-11 7.432572e-20 3.223233e-01
```

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] 7.120929e-11 7.432572e-20 3.223233e-01
```

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! (The `p.adjust()` function automatically takes care of this.)

Table 4: Unadjusted and Bonferroni Adjusted p-Values for Income Comparisons Between Major Categorizations

Comparison	Unadjusted p	Bonferroni-Adjusted p
Science − non-STEM	0.107	0.322
Tech/Eng/Math − non-STEM	<0.001	<0.001
Tech/Eng/Math − Science	<0.001	<0.001

Other p-Value Adjustment Methods

There is nothing that requires you to evenly adjust the p-value across the three 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] 3.560464e-11 7.432572e-20 1.074411e-01
```

Table 5: Unadjusted and Benjamini–Hochberg Adjusted p-Values for Income Comparisons Between Major Categorizations

Comparison	Unadjusted p	Bonferroni-Adjusted p
Science − non-STEM	0.107	0.107
Tech/Eng/Math − non-STEM	<0.001	<0.001
Tech/Eng/Math − Science	<0.001	<0.001

Using the Benjamini–Hochberg adjusted p-values, we still find statistically significant income differences between (1) non-STEM and science majors, and (2) non-STEM and Tech/Eng./Math 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 Major Categorization Mediate the Relationship Between Proportion of Women and Income?

According to Wikipedia,

[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.

> [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"- and non-STEM majors and it is the type of major that is causing the lower incomes. In this sense, we would argue that type of major 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 major categorization 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 + science + tech_eng_math, data = stem)
glance(lm.mediator)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1 0.5018268     0.4929308 8.161516   56.4107 2.813648e-25  4 -603.1358
       AIC      BIC deviance df.residual
1 1216.272 1232.009 11190.54         168
```

At the model-level, we find that differences in the proportion of women who graduate from the major and major categorization explain 50.2% of the variation in median incomes. This is a statistically significant amount of variation, $F(4, 168) = 56.41$, $p < .0001$.

```
tidy(lm.mediator)
```

```
           term   estimate  std.error statistic      p.value
1   (Intercept) 46.2135671 2.21933337 20.823175 2.101870e-48
2         women -0.1851608 0.03488069 -5.308404 3.459773e-07
3       science  3.6136674 1.60272737  2.254699 2.544385e-02
4 tech_eng_math 11.9383312 1.92109975  6.214321 3.928573e-09
```

At the coefficient-level, we are most interested in the β-coefficient associated with the `women` predictor. This coefficient, $\hat{\beta} = -0.19$, indicates that after controlling for type of 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 $185 decrease in median income, on average.

Although this effect is still statistically significant, $t(168) = -5.31$, $p < .001$, the size of the effect has dimished from what it was in the simple regression model ($\hat{\beta} = -0.30$). After controlling for type of major, the effect has diminished. This seems to support our mediation hypothesis that the type of 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, in Hypothesis 3, our focal research was on the income differences between major categorizations. 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.

We will examine whether income difference across major 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.7 = lm(income ~ 1 + women + science + tech_eng_math, data = stem)
lm.8 = lm(income ~ 1 + women + non_stem + tech_eng_math, data = stem)

# Adjust p-values
p_values = c(
  0.02544385, #non-STEM vs. science
  0.00000003928573, #non-STEM vs. Tech/Eng/Math
  0.0003641043 #science vs. Tech/Eng/Math
  )
p.adjust(p_values, method = "BH")
```

```
[1] 0.0254438500000 0.0000001178572 0.0005461564500
```

Table 6: Adjusted Mean Differences, Unadjusted and Benjamini–Hochberg Adjusted p-Values for Income Comparisons Between Major Categories Controlling for Differences in the Percentage of Female Graduates

Comparison	Adj. Mean Difference	Unadjusted p	Adjusted p
Science − non-STEM	$3,613	0.02544	0.0254400
Tech/Eng/Math − non-STEM	$11,938	0.00000	0.0000001
Tech/Eng/Math − Science	$8,325	0.00036	0.0005500

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) non-STEM and Science majors, (2) non-STEM and Tech/Eng./Math majors, and (3) Science and Tech/Eng./Math majors. After controlling for the percentage of women in the major, the income differences between non-STEM and Science majors has become statistically significant!

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 non-STEM and Science majors is $3,613 (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

Comparison	Unadjusted	Adjusted
Science − non-STEM	\$2,781	\$3,613
Tech/Eng/Math − non-STEM	\$18,064	\$11,938
Tech/Eng/Math − Science	\$15,283	\$8,325

The results show how the income difference between major 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 major 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 major 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 data, to test whether there is an effect of major category on income, we evaluate three hypothesis tests:

$$
\begin{aligned}
H_0 :& \mu_{\text{Science}} - \mu_{\text{non-STEM}} = 0 \\
H_0 :& \mu_{\text{Tech-Eng-Math}} - \mu_{\text{non-STEM}} = 0 \\
H_0 :& \mu_{\text{Tech-Eng-Math}} - \mu_{\text{Science}} = 0
\end{aligned}
$$

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 three tests, or in two of the three tests, or in all three of the three tests. 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

$$P(\text{type I error}) = 1 - (1 - 0.05)^3 = 0.142$$

The probability that we will make at least one Type I error in the six tests is .142 NOT .05!!! 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:

$$0.05 = 1 - (1 - \alpha)^3$$

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.

Carnevale, A. P., Strohl, J., & Melton, M. (2011). *What's it worth?: The economic value of college majors.* Georgetown University Center on Education; the Workforce. Retrieved from http://cew.georgetown.edu/whatsitworth

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.