Polychotomous Categorical Predictors

2019-10-27

Preparation

In this set of notes, we will examine the question of whether family structure affects adolescents' use of illegal substances (cigarettes, alcohol, and marijuana). In particular we will evaluate whether adolescents from intact families use these substances at different rates than adolescents from non-intact families. To do so, we will use the *substance-family.csv* data (see the data codebook). To begin, we will load several libraries and import the data into an object called `family`.

```
# Load libraries
library(broom)
library(corrr)
library(ggridges)
library(tidyverse)

# Import data
family = read_csv(file = "~/Documents/github/epsy-8251/data/substance-family.csv")
head(family)
```

```
# A tibble: 6 x 4
  substance_use family_structure     female   gpa
          <dbl> <chr>                 <dbl> <dbl>
1       -0.129  Two-parent family         1   3.8
2        0.0143 Two-parent family         0   2.5
3       -0.594  Two-parent family         1   2.8
4       -0.439  Single-parent family      0   3.5
5       -0.284  Two-parent family         1   3.3
6       -0.284  Two-parent family         0   2.5
```

Examine and Describe the Marginal Distribution of the Adolescent Substance Use

To begin the analysis, we will explore the outcome variable, `substance_use`. Below we examine both the marginal distribution of this variable, and its distribution conditioned on family structure.

```
# Marginal distribution of substance use
ggplot(data = family, aes(x = substance_use, y = ..density..)) +
  geom_histogram(color = "black", fill = "skyblue") +
  stat_density(geom = "line") +
  theme_bw() +
  xlab("Substance use") +
  ylab("Probability density")

# Substance use conditioned on family structure
ggplot(data = family, aes(x = substance_use, y = family_structure)) +
  geom_density_ridges() +
  theme_bw() +
  ylab("Family structure") +
  xlab("Substance use")
```

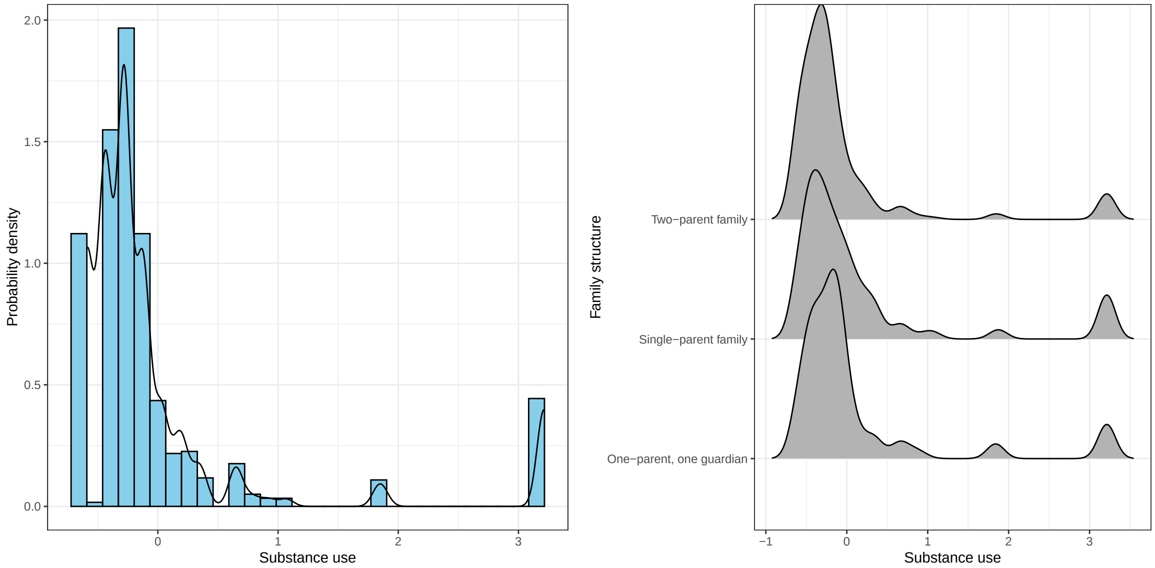

Figure 1. Density plot of the substance use variable.

```
# Compute summary statistics
family %>%
  group_by(family_structure) %>%
  summarize(
    M  = mean(substance_use),
    SD = sd(substance_use)
  )
```

```
# A tibble: 3 x 3
  family_structure               M    SD
  <chr>                      <dbl> <dbl>
1 One-parent, one guardian  0.0991 0.950
2 Single-parent family      0.156  1.04 
3 Two-parent family        -0.0653 0.834
```

The distribution of substance use is highly right-skewed with the majority of adolescents at the lower end of the distribution. After conditioning on family structure, the data suggest that adolescents from two-parent households have a lower mean substance use than other adolescents. Similarly, adolescents from households with one parent and one guardian have a lower mean substance use than adolescents from single-parent households. There is a great deal of variation in all three distributions and adolescents who use substances at a high rate in all three distributions.

Does Family Structure Predict Adolescent Substance Use?

To examine whether the observed difference in substance use between adolescents from the three family structures is more than we would expect because of chance, we can fit a regression model using family structure to predict variation in substance use. Before fitting this model, however, we need to create a dummy variable for EACH category of the `family_structure` variable. For our analysis, we will need to create three dummy variables: `two_parent`, `parent_guardian`, and `one_parent`. 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:

$$\texttt{if\_else}(\text{conditional statement},\ \text{output if TRUE},\ \text{output if FALSE})$$

For example, to create the dummy variable `two_parent`, we examine the `family_structure` column and give the dummy variable a value of `1` if the label is `Two-parent family` (a `TRUE` element in our logical vector) and a value of `0` if it isn't (a `FALSE` element in our logical vector). The full `if_else()` syntax to create a `science` dummy-coded variable is this:

```r
# Create science dummy variable
family %>%
  mutate(
    two_parent = if_else(family_structure == "Two-parent family", 1, 0)
    )
```

```
# A tibble: 910 x 5
   substance_use family_structure     female   gpa two_parent
           <dbl> <chr>                 <dbl> <dbl>      <dbl>
 1       -0.129  Two-parent family         1   3.8          1
 2        0.0143 Two-parent family         0   2.5          1
 3       -0.594  Two-parent family         1   2.8          1
 4       -0.439  Single-parent family      0   3.5          0
 5       -0.284  Two-parent family         1   3.3          1
 6       -0.284  Two-parent family         0   2.5          1
 7       -0.594  Two-parent family         1   2.3          1
 8       -0.284  Two-parent family         1   2.5          1
 9        3.21   Two-parent family         0   3            1
10       -0.594  Two-parent family         0   3            1
# ... with 900 more rows
```

Below we write the syntax to create all three dummy variables and save the new columns in the object `family`.

```r
# Create all three dummy variables
family = family %>%
  mutate(
    two_parent = if_else(family_structure == "Two-parent family", 1, 0),
    parent_guardian = if_else(family_structure == "One-parent, one guardian", 1, 0),
    one_parent = if_else(family_structure == "Single-parent family", 1, 0),
    )

# Examine data
head(family)
```

```
# A tibble: 6 x 7
  substance_use family_structure female   gpa two_parent parent_guardian
          <dbl> <chr>             <dbl> <dbl>      <dbl>           <dbl>
1       -0.129  Two-parent fami~      1   3.8          1               0
2        0.0143 Two-parent fami~      0   2.5          1               0
3       -0.594  Two-parent fami~      1   2.8          1               0
4       -0.439  Single-parent f~      0   3.5          0               0
5       -0.284  Two-parent fami~      1   3.3          1               0
6       -0.284  Two-parent fami~      0   2.5          1               0
# ... with 1 more variable: one_parent <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(family$family_structure)
```

```
[1] "Two-parent family"        "Single-parent family"
[3] "One-parent, one guardian"
```

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 `parent_guardian`, and `one_parent` as predictors in the model; we did not include the `two_parent` predictor. As such, adolescents from two-parent households is our reference group.

```
# Two-parent households is reference group
lm.1 = lm(substance_use ~ 1 + parent_guardian + one_parent, data = family)

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

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1    0.0107       0.00849 0.887      4.89 0.00771     3 -1180. 2369. 2388.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

At the model-level, differences in family structure explain 1.06% of the variation in adolescents' substance use. The empirical data are not consistent with the hypothesis that family structure does not explain variation in adolescents' substance use, $F(2, 907) = 4.89$, $p = .008$.

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

```
# A tibble: 3 x 5
  term            estimate std.error statistic p.value
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)      -0.0653    0.0350     -1.87 0.0623
2 parent_guardian   0.164     0.0933      1.76 0.0785
3 one_parent        0.221     0.0778      2.84 0.00455
```

The fitted regression equation is

$$\hat{\text{Substance Use}_i} = -0.07 + 0.16(\text{Parent/Guardian}_i) + 0.22(\text{One-Parent}_i)$$

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 substance use for adolescents from two-parent households is -0.07.
- Adolescents from housholds with one parent and one guardian use substances at a rate that is 0.16 higher on average than adolescents from two-parent households.
- Adolescents from one parent housholds use substances at a rate that is 0.22 higher on average than adolescents from two-parent households.

The statistical hypothesis associated with each of the parameters in the model are:

- $H_0 : \beta_0 = 0$
- $H_0 : \beta_{\text{Parent/Guardian}} = 0$
- $H_0 : \beta_{\text{One-Parent}} = 0$

These relate to the following scientific hypotheses:

- The average substance use for adolescents from two-parent households (reference group) is 0.
- The average substance use for adolescents housholds with one parent and one guardian is not different than the average substance use for adolescents from two-parent households.
- The average substance use for adolescents from one parent households is not different than the average substance use for adolescents from two-parent households.

Because the scientific hypotheses are really about comparisons of conditional means, the statistical hypotheses can also be written to reflect this as:

- $H_0 : \mu_{\text{Two-Parent}} = 0$
- $H_0 : \mu_{\text{Two-Parent}} = \mu_{\text{Parent/Guardian}}$ or equivalently $H_0 : \mu_{\text{Parent/Guardian}} - \mu_{\text{Two-Parent}} = 0$
- $H_0 : \mu_{\text{Two-Parent}} = \mu_{\text{One-Parent}}$ or equivalently $H_0 : \mu_{\text{One-Parent}} - \mu_{\text{Two-Parent}} = 0$

It is evaluation of the latter two hypotheses (those associated with the partial slopes in the model) that allow us to answer our research question of whether adolescents from intact families have lower rates of substance use than other adolescents. The *p*-values associated with the two partial slope coefficients indicate that the data are not consistent with the hypothesis of no difference in the rates of substance use between the reference group (two-parent households) and the family structure identified in each of the dummy coded predcitors. This implies that it is likely that the rates of substance use for both adolescents from housholds with one parent and one guardian ($t_{907} = 1.76$, $p = .008$) and those from one parent housholds ($t_{907} = 2.84$, $p = .005$) are higher than the substance use rate for adolescents from two-parent households.

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:

$$\beta_{\text{Parent/Guardian}} = \mu_{\text{Parent/Guardian}} - \mu_{\text{Two-Parent}}$$
$$\beta_{\text{One Parent}} = \mu_{\text{One-Parent}} - \mu_{\text{Two-Parent}}$$

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. In general,

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

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 three sets of pairwise diffeences:

- $\mu_{\text{Parent/Guardian}} - \mu_{\text{Two-Parent}}$
- $\mu_{\text{One-Parent}} - \mu_{\text{Two-Parent}}$
- $\mu_{\text{One-Parent}} - \mu_{\text{Parent/Guardian}}$

Thus we can express the model-level null hypothesis using partial effects as:

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

or using mean differences as:

$$\begin{aligned} H_0 : &\bigg(\mu_{\text{Parent/Guardian}} - \mu_{\text{Two-Parent}}\bigg) = \bigg(\mu_{\text{One-Parent}} - \mu_{\text{Two-Parent}}\bigg) = \\ &\bigg(\mu_{\text{One-Parent}} - \mu_{\text{Parent/Guardian}}\bigg) = 0 \end{aligned}$$

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.

Evaluate Substance Use Bewteen Adolescents from Single-Parent Households and Those from Houeholds with One Parent and One Guardian

In order to examine the remaining pairwise difference, we need to fit an additional regression model that allows us to evaluate this last comparison between adolescents from single-parent households and those from houeholds with one parent and one guardian. Below, we fit a second model (using single-parent households as the reference group) to predict variation in substance use

```r
# Single-parent households is reference group
lm.2 = lm(substance_use ~ 1 + parent_guardian + two_parent, data = family)

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

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1    0.0107       0.00849 0.887      4.89 0.00771     3 -1180. 2369. 2388.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Note that the model-level output for this fitted model is exactly the same as that for the model in which two-parent households was the reference group. This is because we are fitting the exact same omnibus model (to examine whether the three sets of pairwise differences explain variation in substance use).

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

```
# A tibble: 3 x 5
  term            estimate std.error statistic p.value
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)       0.156     0.0694     2.24  0.0250
2 parent_guardian  -0.0568    0.111     -0.512 0.609
3 two_parent       -0.221     0.0778    -2.84  0.00455
```

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

$$\widehat{\text{Substance Use}}_i = 0.16 - 0.06(\text{Parent/Guardian}_i) - 0.22(\text{Two-Parent}_i)$$

Interpreting these values,

- The average substance use for adolescents from single-parent households is 0.16. The empirical evidence suggests that this rate of substance use is statistically different than 0; $t_{907} = 2.25$, $p = .025$.
- Adolescents from housholds with one parent and one guardian use substances at a rate that is 0.06 lower on average than adolescents from single-parent households. The empirical evidence suggests that this rate of substance use is not more than we would expect because of sampling variation; $t_{907} = -0.51$, $p = .609$.
- Adolescents from two-parent housholds use substances at a rate that is 0.22 lower on average than adolescents from two-parent households. The empirical evidence suggests that this difference in the rate of substance use is more than we would expect because of sampling variation; $t_{907} = -2.84$, $p = .005$.

Link to the Analysis of Variance Methodology for Testing Mean Differences

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:

$$H_0 : \mu_{\text{Two-Parent}} = \mu_{\text{Parent/Guardian}} = \mu_{\text{One-Parent}}$$

This is the way we write the omnibus null hypothesis that is associated with the one-factor analysis of variance (ANOVA). Fitting a regression model with dummy-variables gives the exact same results 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 ($H_0 : \mu_{\text{Two-Parent}} = \mu_{\text{Parent/Guardian}} = \mu_{\text{One-Parent}}$); these are just different ways of writing the model-level null hypothesis!

Further Understanding Differences in Substance Use

If you are only interested in if there are differences in adolescent substance use between the three family structures, you can focus on the model-level (omnibus) results. If, however, you want to go further and understand the nature of those differences, in particular whether the adolescent substance use for each family structure differs from the others, it is necessary to examine the **pairwise differences** between family structures. Based on the two sets of coefficeint-level results, the pairwise differences are:

Table 1
Pairwise Comparisons of Adolescent Substance Use between Three Family Structures

Comparison	Mean Difference	p
Two-parent vs. Parent/guardian	0.164	0.079
Two-parent vs. Single-parent	0.221	0.005
Parent/guardian vs. Single-parent	0.057	0.609

Multiple Comparisons

When we evaluated the p-values for each of these pairwise differences, we used an unadjusted *p*-value (the *p*-value from the `tidy()` output). 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 categorical predictor with more than two levels, however, the null hypothesis constitutes more than one mean difference.

Remember, the effect of family structure constitutes 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 adjust the *p*-value to compensate for this increased number of effects/mean differences inherent in the family structure effect. There are many ways to compensate for this increased number of mean differences in the effect, and all of them adjust the *p*-value of each mean difference assoicated with family structure.

Bonferonni Adjusted p-Values

The easeist way to make these adjustments is to multiply each *p*-value associated with the effect of family structure by 3 (the number of mean differences associated with family structure).

$$p_{\text{Adjusted}} = p \times 3$$

This is referred to as the Dunn-Bonferonni adjustment, named for Olvie Dunn and Carlo Bonferroni. In general, the adjustment is

$$p_{\text{Adjusted}} = p \times k,$$

where k is the number of pairwise differences encompassed in the effect. For our example,

```r
c(.079,.005,.609) * 3
```

```
[1] 0.237 0.015 1.827
```

In practice, since *p*-values have limits of 0 and 1, any adjusted *p*-value that exceeds 1 is limited to 1.

Table 2

Pairwise Comparisons of Adolescent Substance Use between Three Family Structures

Comparison	Mean Difference	p	Bonferroni Adjusted p
Two-parent vs. Parent/guardian	0.164	0.079	0.237
Two-parent vs. Single-parent	0.221	0.005	0.015
Parent/guardian vs. Single-parent	0.057	0.609	1.000

After adjusting the *p*-values, the difference we saw earlier in the average substance use between adolescents from two-parent housholds and those from households with one parent and one guardian, is now gone. The empirical evidence now suggests there is not a difference between these two groups. In other words, what we saw earlier was likely a type I error. The whole goal of *p*-value adjustment is to protect against type I errors!

We can adjust the *p*-values using the Bonferroni adjustment directly with the `p.adjust()` function. To do this, we initially create a vector of the unadjusted *p*-values using the `c()` function and then include this vector in the `p.adjust()` function along with the argument `method = "bonferroni"`.

```
# Create vector of unadjusted p-values
p_values = c(0.079, 0.005, 0.609)

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

```
[1] 0.237 0.015 1.000
```

Note that the `p.adjust()` function automatically sets the upper-limit of the reported *p*-values to 1.

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 adjusted p-values
p.adjust(p_values, method = "BH")
```

```
[1] 0.1185 0.0150 0.6090
```

[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.

Table 3

Pairwise Comparisons of Adolescent Substance Use between Three Family Structures

Comparison	Mean Difference	p	Bonferroni Adjusted p	Benjamini–Hochberg p
Two-parent vs. Parent/guardian	0.164	0.079	0.237	0.119
Two-parent vs. Single-parent	0.221	0.005	0.015	0.015
Parent/guardian vs. Single-parent	0.057	0.609	1.000	0.609

After adjusting the *p*-values using the Benjamini–Hochberg method, the empirical evidence suggests that only adolescents from two-parent housholds and those from single-parent households differ in their average substance use. While this is the same general result we found when we used the Bonferroni-adjusted *p*-values, the comparison between adolescents from two-parent housholds and those from households with one parent and one guardian has a much lower *p*-value using the Benjamini–Hochburg adjustment than using the Bonferroni adjustment. This suggests that the Benjamini–Hochburg adjustment provides more statistical power to find group differences than the Bonferroni adjustment.

It is importnant to note that the Benjamini–Hochburg adjustment does not protect as well against type I error as the Bonferonni adjustment does. Instead, the Benjamini–Hochburg adjustment protects against **false discovery**. In the research community, the current thinking is thatit is more beneficial to protect against false discovery than type I error, especially in exploratory research.

Which Adjustment Method Should I Use?

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 Family Structure Predict Adolescent Substance Use After Accounting for Other Covariates?

One question we may have is whether the differences we saw in adolescents' average substance use persist after we account for other covariates that also explain differences in substance use (e.g., sex, academic achievement). To evaluate this, we will fit a regression model that includes the `female` and `gpa` covariates along with two of the dummy-coded family structure predictors to explain variability in substance use.

```r
# Two-parent households is reference group
lm.3 = lm(substance_use ~ 1 + female + gpa + parent_guardian + one_parent, data = family)

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

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1    0.0226        0.0183 0.882      5.24 3.56e-4     5 -1175. 2361. 2390.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

At the model-level, differences in family structure, sex of the adolescent, and composite GPA explain 2.26% of the variation in adolescents' substance use. The empirical data are not consis5.24\$, $p < .001$.

```r
tidy(lm.3)
```

```
# A tibble: 5 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      0.0170     0.0945     0.180 0.857
2 female          -0.196      0.0588    -3.33  0.000916
3 gpa              0.00496    0.0268     0.185 0.853
4 parent_guardian  0.158      0.0936     1.69  0.0920
5 one_parent       0.210      0.0776     2.71  0.00682
```

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

$$\widehat{\text{Substance Use}}_i = 0.02 - 0.20(\text{Female}_i) + 0.005(\text{GPA}_i) + 0.16(\text{Parent/Guardian}_i) + 0.22(\text{One-Parent}_i)$$

To answer our research question, we are most interested in the partial regression coefficients associated with the family structure variables. Interpreting these values,

- Adolescents from housholds with one parent and one guardian use substances at a rate that is 0.16 higher on average than adolescents from two-parent households, after controlling for differences in adolescents' sex and GPA.
- Adolescents from one-parent housholds use substances at a rate that is 0.22 higher on average than adolescents from two-parent households.

To determine whether there are differences in the average substance use between adolescents from one-parent households and those from households with one parent and one guardian, we need to fit an additional model with one of these groups as the reference group.

```r
# One-parent households is reference group
lm.4 = lm(substance_use ~ 1 + female + gpa + parent_guardian + two_parent, data = family)

tidy(lm.4)
```

```
# A tibble: 5 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      0.227      0.108      2.11  0.0347
2 female          -0.196      0.0588    -3.33  0.000916
3 gpa              0.00496    0.0268     0.185 0.853
4 parent_guardian -0.0525     0.111     -0.475 0.635
5 two_parent      -0.210      0.0776    -2.71  0.00682
```

- Adolescents from housholds with one parent and one guardian use substances at a rate that is 0.05 lower on average than adolescents from one-parent households, after controlling for differences in adolescents' sex and GPA.

To evaluate whether these differences are more thanwe expect because of chance (sampling variation), we need to again adjust the *p*-values because we are evaluating three mean differences. We will do this using the Benjamini–Hochberg adjustment.

```r
# Create vector of unadjusted p-values
p_values = c(0.092, 0.007, 0.092)

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

```
[1] 0.092 0.021 0.092
```

Based on the Benjamini–Hochberg adjusted *p*-values, the only statistically relevant difference in adolescents' substance use if between those living in housholds with two-parents and those living with one-parent. The empirical evidence suggests that this difference in the rate of substance use is more than we would expect because of sampling variation; $t_{905} = 2.71$, Benjamini–Hochberg adjusted $p = .021$.

The empirical evidence for the other two comparisosn are far more uncertain. Both of these differences are consistent with differences that are only due to sampling error. However, after controlling for differences in adolescents' sex and GPA, the *p*-values for both differences are much smaller than in the uncontrolled models. This might be worth noting in any write-up of the analyses.

More importantly, the effects representing the mean differences (the estimated coefficients), did not change much in magnitude nor direction. This suggests that the differences in adolescent substance use between the three family structures is stable, at least when we control for sex and GPA. Moreover, the uncertainty (as measured by the SEs) for these effects also remained stable from the uncontrolled to the controlled model.

Controlled Mean Differences

In the language of Analysis of Covariance (ANCOVA), the controlled mean differences are referred to as *Adjusted Mean Differences*. So, for example, the adjusted mean difference in substance use between adolescents living in two-aprent households and those living in one-parent households is 0.22 (controlling for differences in the adolescents' sex and GPA). 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 4
Unadjusted and Adjusted Mean Differences for Adolescent Substance between Three Family Structures

	Mean Difference	
Comparison	Unadjusted	Adjusted
Two-parent vs. Parent/guardian	0.164	0.158
Two-parent vs. Single-parent	0.221	0.210
Parent/guardian vs. Single-parent	0.057	0.053

Note: The adjusted mean differences were obtained by controlling for adolescent sex and GPA.

Technical Appendix: Type I Error Rate andFalse-Discovery Rate

When we use an alpha value of 0.05 to evaluate consistency of the empirical data to the null hypothesis, we are saying we are willing to make a Type I error in 5% of the samples that could be randomly selected. In other words, we will end up wrongly concluding that the empirical data are inconsistent with the null hypothesis in 5% of the samples we would obtain from our thought experiment. (In practice, 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 using this criterion. 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 family strusture on substance use, we evaluate three hypothesis tests:

- $H_0 : \mu_{\text{Two-Parent}} = 0$
- $H_0 : \mu_{\text{Two-Parent}} = \mu_{\text{Parent/Guardian}}$ or equivalently $H_0 : \mu_{\text{Parent/Guardian}} - \mu_{\text{Two-Parent}} = 0$
- $H_0 : \mu_{\text{Two-Parent}} = \mu_{\text{One-Parent}}$ or equivalently $H_0 : \mu_{\text{One-Parent}} - \mu_{\text{Two-Parent}} = 0$

$$
\begin{aligned}
H_0 :&\mu_{\text{Two-Parent}} - \mu_{\text{Parent/Guardian}} = 0 \\
H_0 :&\mu_{\text{Two-Parent}} - \mu_{\text{One-Parent}} = 0 \\
H_0 :&\mu_{\text{Parent/Guardian}} - \mu_{\text{One-Parent}} = 0
\end{aligned}
$$

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:

$$P(\text{type I error}) = 1 - (1 - \alpha)^k$$

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 0.142 NOT 0.05!!! 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 minimizing *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.

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.