Categorical Independent Variables and Interactions

Dr. Sarah Hunter

3/27/2020

Categorical Independent Variables

Including categorical independent variables into a linear regression model mostly makes interpretation more complicated, not the actual model. However, there can be some problems with R estimating a model with a categorical IV. Using the American National Election Survey 2004 subset, I will demonstrate one issue. First, loading the data:

```
#Setting the working directory
setwd("/Users/sarahhunter/Desktop/Data")

#Loading CSV data
mydata<-read.csv("nes2004subset3.csv")

summary(mydata)
```

```
##        X             religion          bush           female      
##  Min.   :   1.0   Min.   :1.000   Min.   :0.000   Min.   :0.0000  
##  1st Qu.: 302.5   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000  
##  Median : 607.0   Median :1.000   Median :1.000   Median :1.0000  
##  Mean   : 606.1   Mean   :2.311   Mean   :0.508   Mean   :0.5319  
##  3rd Qu.: 908.5   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.0000  
##  Max.   :1212.0   Max.   :7.000   Max.   :1.000   Max.   :1.0000  
##                   NA's   :15      NA's   :396                     
##    unionhouse        partyid         eval_WoT         eval_HoE      
##  Min.   :0.0000   Min.   :0.000   Min.   :-2.000   Min.   :-2.0000  
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:-2.000   1st Qu.:-2.0000  
##  Median :0.0000   Median :3.000   Median : 1.000   Median :-1.0000  
##  Mean   :0.1724   Mean   :2.872   Mean   : 0.152   Mean   :-0.3467  
##  3rd Qu.:0.0000   3rd Qu.:5.000   3rd Qu.: 2.000   3rd Qu.: 2.0000  
##  Max.   :1.0000   Max.   :6.000   Max.   : 2.000   Max.   : 2.0000  
##  NA's   :6        NA's   :17      NA's   :29       NA's   :36       
##     ideology       bush_therm       education         income     
##  Min.   :1.000   Min.   :  0.00   Min.   :0.000   Min.   : 1.00  
##  1st Qu.:3.000   1st Qu.: 30.00   1st Qu.:3.000   1st Qu.:11.00  
##  Median :4.000   Median : 60.00   Median :4.000   Median :16.00  
##  Mean   :4.268   Mean   : 54.94   Mean   :4.305   Mean   :14.97  
##  3rd Qu.:6.000   3rd Qu.: 85.00   3rd Qu.:6.000   3rd Qu.:20.00  
##  Max.   :7.000   Max.   :100.00   Max.   :7.000   Max.   :23.00  
##  NA's   :288                                      NA's   :141    
```

Look at the "religion" variable. We know this is a categorical variable, but it has already been coded with numbers representing various religions. According to the codebook, *1= Protestant* 2= Catholic *4= Jewish* 6= Other *7= None

R reads this as a numeric variable, not a categorical. To show this, we can ask R which `class` of variable religion is:

```
class(mydata$religion)
```

```
## [1] "integer"
```

Because R reads this variable as an integer, if we were to simply put it into a model the usual way, we would get the following:

```
model1<-lm(bush_therm~ education+income+religion, data=mydata)

summary(model1)
```

```
##
## Call:
## lm(formula = bush_therm ~ education + income + religion, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.306 -26.302   4.686  27.300  66.236
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.8188     3.4290  17.736  < 2e-16 ***
## education    -3.0109     0.7043  -4.275 2.08e-05 ***
## income        0.8460     0.1885   4.488 7.97e-06 ***
## religion     -2.5460     0.4657  -5.466 5.73e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.8 on 1053 degrees of freedom
##   (150 observations deleted due to missingness)
## Multiple R-squared:  0.05467,    Adjusted R-squared:  0.05198
## F-statistic:  20.3 on 3 and 1053 DF,  p-value: 8.572e-13
```

In this case, we could read this as "A one unit increase in religion leads to a 2.546 decrease in the respondent's thermometer rating for President Bush". As religion is a categorical variable, that makes no sense. Therefore, we have to tell R to treat religion as a categorical variable with the `as.factor()` command. This command tells R that religion is a factor variable and should be treated as such. The following code shows how to use it.

```
model2<-lm(bush_therm~ education+income+as.factor(religion), data=mydata)

summary(model2)
```

```
##
## Call:
## lm(formula = bush_therm ~ education + income + as.factor(religion),
##     data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.026 -26.246   5.067  27.402  64.497
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           57.3824     3.3795  16.979  < 2e-16 ***
## education             -2.8092     0.7062  -3.978 7.43e-05 ***
## income                 0.9037     0.1892   4.777 2.03e-06 ***
## as.factor(religion)2  -4.4951     2.4388  -1.843  0.06559 .
## as.factor(religion)4 -23.3918     6.1204  -3.822  0.00014 ***
## as.factor(religion)6 -21.9750     8.5821  -2.561  0.01059 *
## as.factor(religion)7 -13.8634     2.8772  -4.818 1.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.71 on 1050 degrees of freedom
##   (150 observations deleted due to missingness)
## Multiple R-squared:  0.06235,    Adjusted R-squared:  0.05699
## F-statistic: 11.64 on 6 and 1050 DF,  p-value: 1.196e-12
```

Now, we can see that we have a separate coefficient estimated for each category of religion, except the omitted category. This omitted category is also called the **reference category**. We then interpret all of our categorical variable coefficients with respect to our reference category. Therefore, Catholic respondents (when religion=2), on average, rate President Bush 4.4951 points *lower* than Protestant respondents (religion=1). Jewish respondents (religion=4), on average, rate President Bush 23.3918 points lower than Protestant respondents.

Changing the Reference Category

Perhaps you did not want Protestant to be the reference category of our model. You can tell R which category to use as the reference category with the `relevel` command. To use that command, you can use one of two strategies. The first is to just include `relevel` into the actual `lm` command. You can do this as:

```
model3<-lm(bush_therm~ education+income+relevel(as.factor(religion), ref=2), data=mydata)

summary(model3)
```

```
##
## Call:
## lm(formula = bush_therm ~ education + income + relevel(as.factor(religion),
##     ref = 2), data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.026 -26.246   5.067  27.402  64.497
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             52.8873     3.8188  13.849  < 2e-16 ***
## education                               -2.8092     0.7062  -3.978 7.43e-05 ***
## income                                   0.9037     0.1892   4.777 2.03e-06 ***
## relevel(as.factor(religion), ref = 2)1   4.4951     2.4388   1.843  0.06559 .
## relevel(as.factor(religion), ref = 2)4 -18.8967     6.2793  -3.009  0.00268 **
## relevel(as.factor(religion), ref = 2)6 -17.4799     8.7047  -2.008  0.04489 *
## relevel(as.factor(religion), ref = 2)7  -9.3682     3.2522  -2.881  0.00405 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.71 on 1050 degrees of freedom
##   (150 observations deleted due to missingness)
## Multiple R-squared:  0.06235,    Adjusted R-squared:  0.05699
## F-statistic: 11.64 on 6 and 1050 DF,  p-value: 1.196e-12
```

Model 3 uses Catholic as the reference category. You can do this for any of the religion categories.

The second method of changing the reference category is to change the default reference category in the dataset. You can do this with the following code:

```
mydata<-within(mydata, religion<-relevel(as.factor(religion), ref= 4))

model4<-lm(bush_therm~ education+income+religion, data=mydata)
summary(model4)
```

```
##
## Call:
## lm(formula = bush_therm ~ education + income + religion, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.026 -26.246   5.067  27.402  64.497
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.4074     9.2420   3.831 0.000135 ***
## education    -2.8092     0.7062  -3.978 7.43e-05 ***
## income        0.9037     0.1892   4.777 2.03e-06 ***
## religion1    21.9750     8.5821   2.561 0.010589 *
## religion2    17.4799     8.7047   2.008 0.044889 *
## religion4    -1.4168    10.2986  -0.138 0.890605
## religion7     8.1117     8.8413   0.917 0.359105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.71 on 1050 degrees of freedom
##   (150 observations deleted due to missingness)
## Multiple R-squared:  0.06235,    Adjusted R-squared:  0.05699
## F-statistic: 11.64 on 6 and 1050 DF,  p-value: 1.196e-12
```

Interactive Models

Occasionally, we hypothesize that the effect of one variable depends on the value of another variable. In this case, we would need an **interactive model**, which is a model that includes an interactive term. An interactive term is created by multiplying two independent variables together. Therefore, our new regression model looks like:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 * X_i + \hat{\beta}_2 * Z_i + \hat{\beta}_3 * (Z_i * X_i) + \hat{u}_i$$

For example, we could hypothesize that the effect of income on how survey respondents rate President George W. Bush on the Feeling Thermometer scale depends on if the respondent's household has union membership. Then we can plug in our variables into the general regression equation from above:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 * income_i + \hat{\beta}_2 * unionhouse_i + \hat{\beta}_3 * (unionhouse_i * income_i) + \hat{u}_i$$

We use interactive terms if we think that union respondents have a different slope parameter for income (a different effect of income on Bush's feeling thermometer score). We can demonstrate the difference by plugging in the values (0 and 1) for the unionhouse variable. Then we get two different outcomes: one for union respondents and one for the rest:

For unionhouse = 1:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 * income_i + \hat{\beta}_2 * 1 + \hat{\beta}_3 * (1 * income_i) + \hat{u}_i$$

which simplifies (by collecting terms) to:

$$Y_i = (\hat{\alpha} + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) * income_i + \hat{u}_i$$

For unionhouse=0

$$Y_i = \hat{\alpha} + \hat{\beta}_1 * income_i + \hat{\beta}_2 * 0 + \hat{\beta}_3 * (0 * income_i) + \hat{u}_i$$

which simplifies (by collecting terms) to:

$$Y_i = (\hat{\alpha}) + \hat{\beta}_1 * income_i + \hat{u}_i$$

Therefore, we can see that the effect of income for a union member respondent's household is $\hat{\beta}_1 + \hat{\beta}_3$ while the effect of income for non-union respondents is $\hat{\beta}_1$.

You can also interact two continuous variables. For example, if we think that the effect of income on the respondent's Bush feeling thermometer score depends on the level of education, we could write our regression equation:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 * income_i + \hat{\beta}_2 * education_i + \hat{\beta}_3 * (education_i * income_i) + \hat{u}_i$$

However, the math becomes har more difficult and you need partial derivatives to get the slope coefficient for just education or just income. Because of this, I will skip the math of continuous interactions and move to the R code behind interactive models.

Interactive Models in R

Implementing interactive models in R is really a small twist on the familiar linear regression model in R. Using the same data as above, we will go through the steps of first an interactive model with a categorical variable and then an interactive model with two continuous variables.

Interactive Models with Categorical Variables

To include and interactive term in a linear model in R, you simply need to add the `*` character between two variables:

Note: sometimes, even with dummy variables, you need to specify that `unionhouse` is a factor variable, not numeric. Use the `as.factor()` command for this.

```
mydata$unionhouse<-as.factor(mydata$unionhouse)

interactive.model1<-lm(bush_therm~ female+ income + unionhouse*income, data=mydata)

summary(interactive.model1)
```

```
##
## Call:
## lm(formula = bush_therm ~ female + income + unionhouse * income,
##     data = mydata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.70 -28.25   6.21  26.98  56.22
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            48.0206     3.2026  14.994  < 2e-16 ***
## female                 -3.1844     2.0654  -1.542 0.123425    
## income                  0.6819     0.1850   3.687 0.000239 ***
## unionhouse1            -0.8989     9.7002  -0.093 0.926182    
## income:unionhouse1     -0.6925     0.5517  -1.255 0.209711    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 33.27 on 1056 degrees of freedom
##   (146 observations deleted due to missingness)
## Multiple R-squared:  0.03111,    Adjusted R-squared:  0.02744 
## F-statistic: 8.477 on 4 and 1056 DF,  p-value: 9.854e-07
```

These results look familiar. The results presented here look just like the normal results from a linear regression. The main difference is the last term (`income:unionhouse1`). This is the $\hat{\beta}_3$ from the example above.

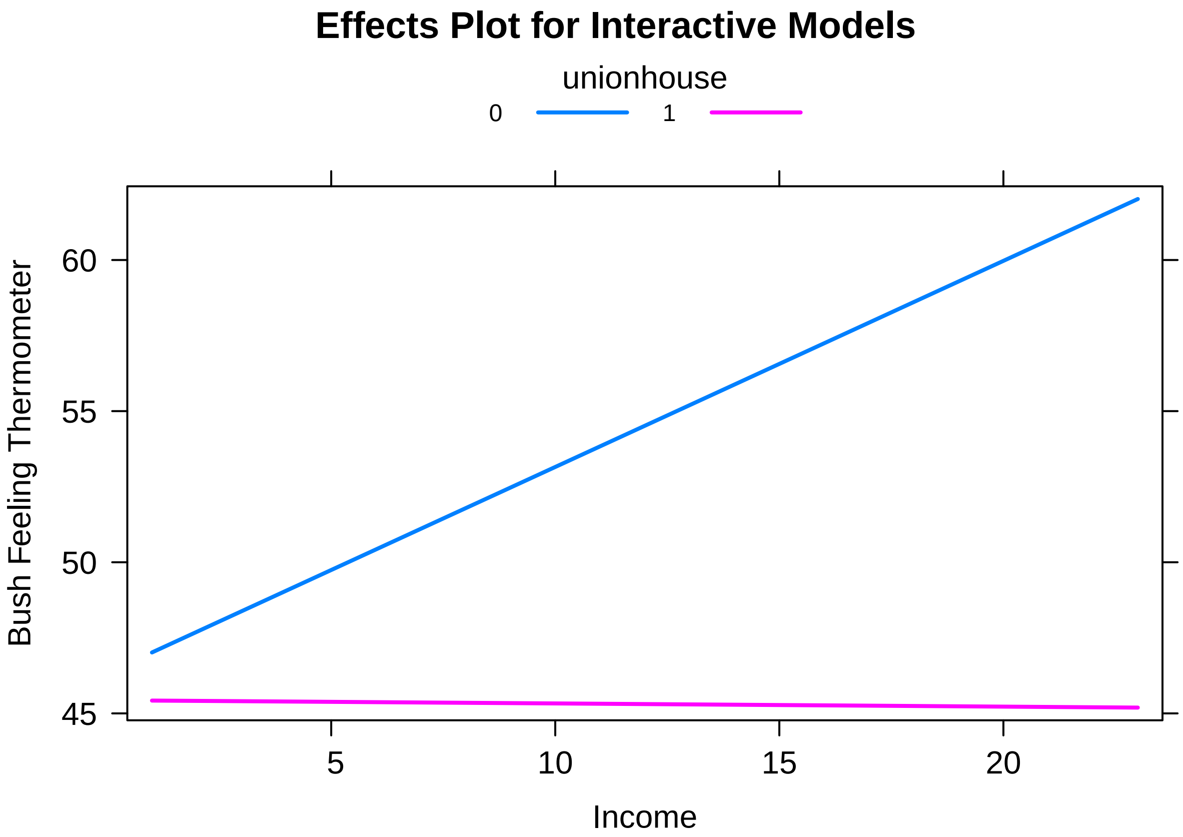
Generally speaking, the best way to present the results from an interactive model is using plots. The `effects` package used in the regression file would also work well here. Below is an example of how to plot the effect of income on the Bush feeling thermometer score, conditional on whether or not the respondent has a household member with union membership:

```r
library(effects)
```

```
## Loading required package: carData
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

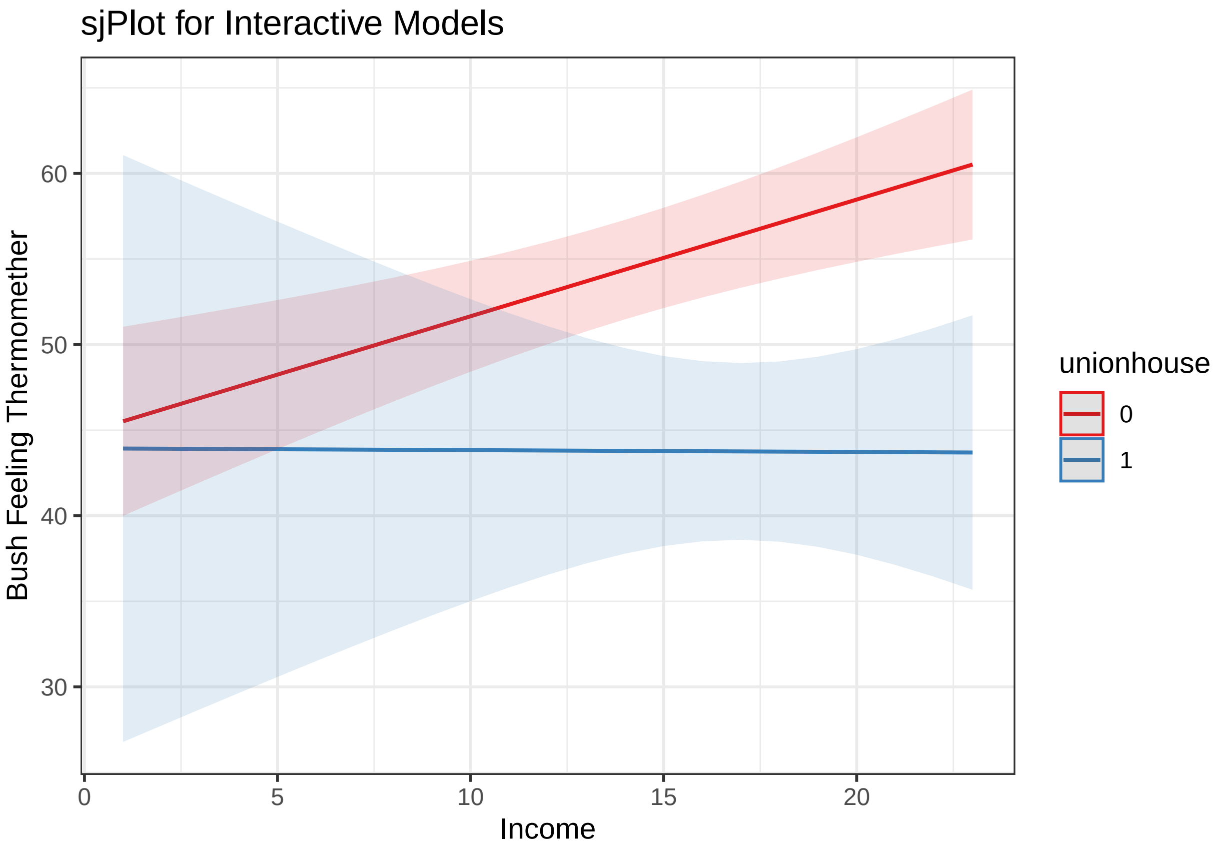
```r
plot(effect("income*unionhouse", interactive.model1), x.var="income", z.var="unionhouse",
     multiline=TRUE, rug=FALSE, xlab="Income", ylab="Bush Feeling Thermometer",
     main="Effects Plot for Interactive Models")
```



With this plot, you have to first specify the term you want to plot (the interactive term), the model (the name of the interactive model), the x variable of which you want to know the effect on the dependent variable (income in this case), and the z variable that moderates the x variable (in this case, union membership). The option `multiline=TRUE` tells the effects plot to put it all on one plot, rather than making two plots. As you can see, the slopes for income are different dependent on whether the respondent was part of a union or non-union household. This example demonstrates: 1. why interactive models are useful and 2. why plots are important parts of presenting results of interactive models.

You can also use a different package called `sjPlot` to get a different looking plot for an interactive model:

```
#install.packages("sjPlot")
#install.packages("ggplot2")
library(sjPlot)
library(ggplot2)

plot_model(interactive.model1, type="pred", terms=c("income", "unionhouse"))+
  xlab("Income")+ylab("Bush Feeling Thermomether")+
  ggtitle("sjPlot for Interactive Models")+theme_bw()
```



`sjPlot` uses a syntax called "ggplot", which is a sophisticated plotting package in R. As you can see, it produces some pretty cool plots, but does take time to learn.

Interactive Models with Continuous Variables

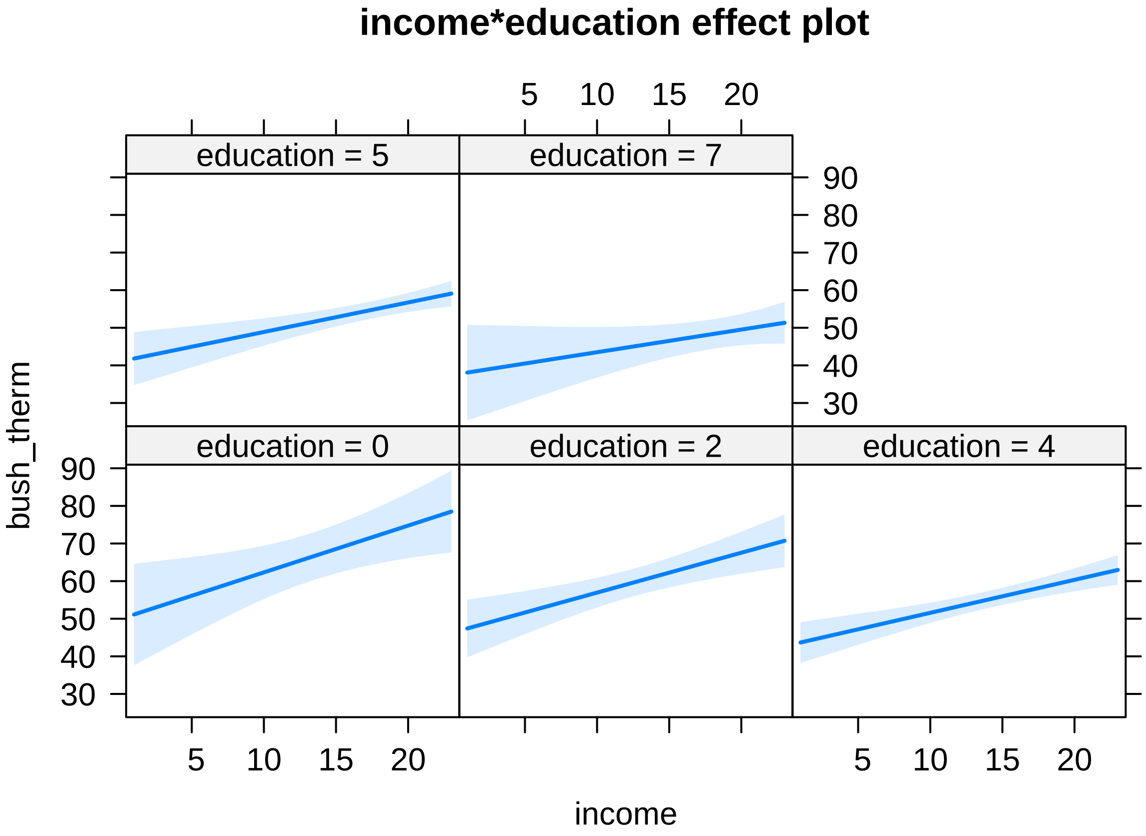
As mentioned previously, we can also interact two continuous variables. The actual syntax in R does not change. We use the same code for interactions with continuous variable as with a categorical variable. Below is an example:

```
interactive.model2<-lm(bush_therm~ income +education +income*education, data=mydata)

summary(interactive.model2)
```

```
##
## Call:
## lm(formula = bush_therm ~ income + education + income * education,
##     data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.549 -26.846   4.857  28.242  58.658
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      49.88870    7.29457   6.839 1.34e-11 ***
## income            1.24371    0.48203   2.580    0.010 *
## education        -1.77116    1.84198  -0.962    0.336
## income:education -0.09174    0.10877  -0.843    0.399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.31 on 1062 degrees of freedom
##   (141 observations deleted due to missingness)
## Multiple R-squared:  0.02717,    Adjusted R-squared:  0.02442
## F-statistic: 9.886 on 3 and 1062 DF,  p-value: 1.969e-06
```

As I mentioned earlier, the interpretation of these coefficients is slightly more complicated. While we could write out the exact slope for union and non-union households, we cannot do the same for an interaction term with two continuous variables. Plots are the only way forward. However, these plots will also look different. Some plots will look at the effect of income at various levels of education, such as with the `effects` package.

```r
plot(effect("income*education", interactive.model2), x.var="income", z.var="education",
     rug=FALSE)
```

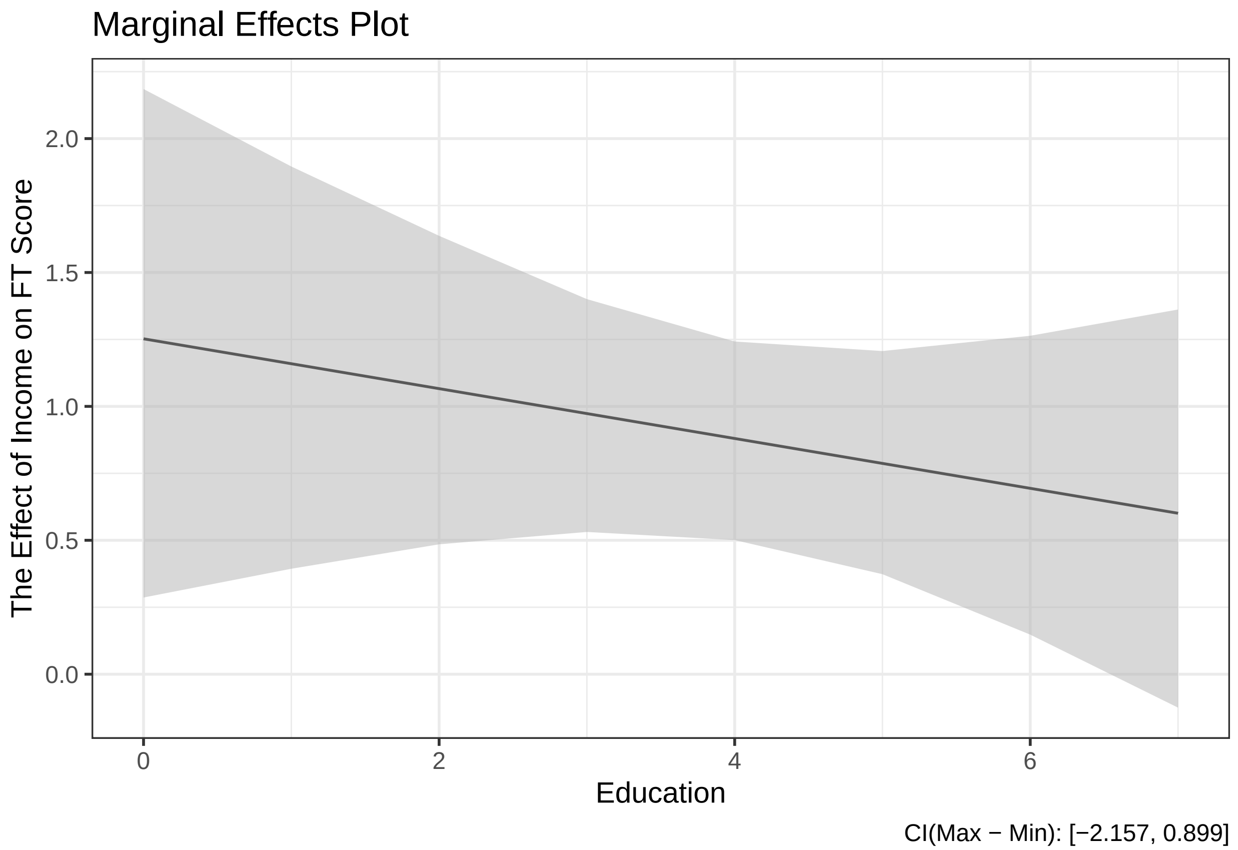


We could also make what are called "marginal effect" plots. These show the actual value of the coefficient for one variable at each level of the other variable, which is what the `interplot` package gives:

```r
#if you need to install:

#install.packages("interplot")

#always:
library(interplot)
```

```
## Loading required package: abind

## Loading required package: arm

## Loading required package: MASS

## Loading required package: Matrix

## Loading required package: lme4

##
## arm (Version 1.10-1, built: 2018-4-12)

## Working directory is /Users/sarahhunter/Desktop/R Markdown
```

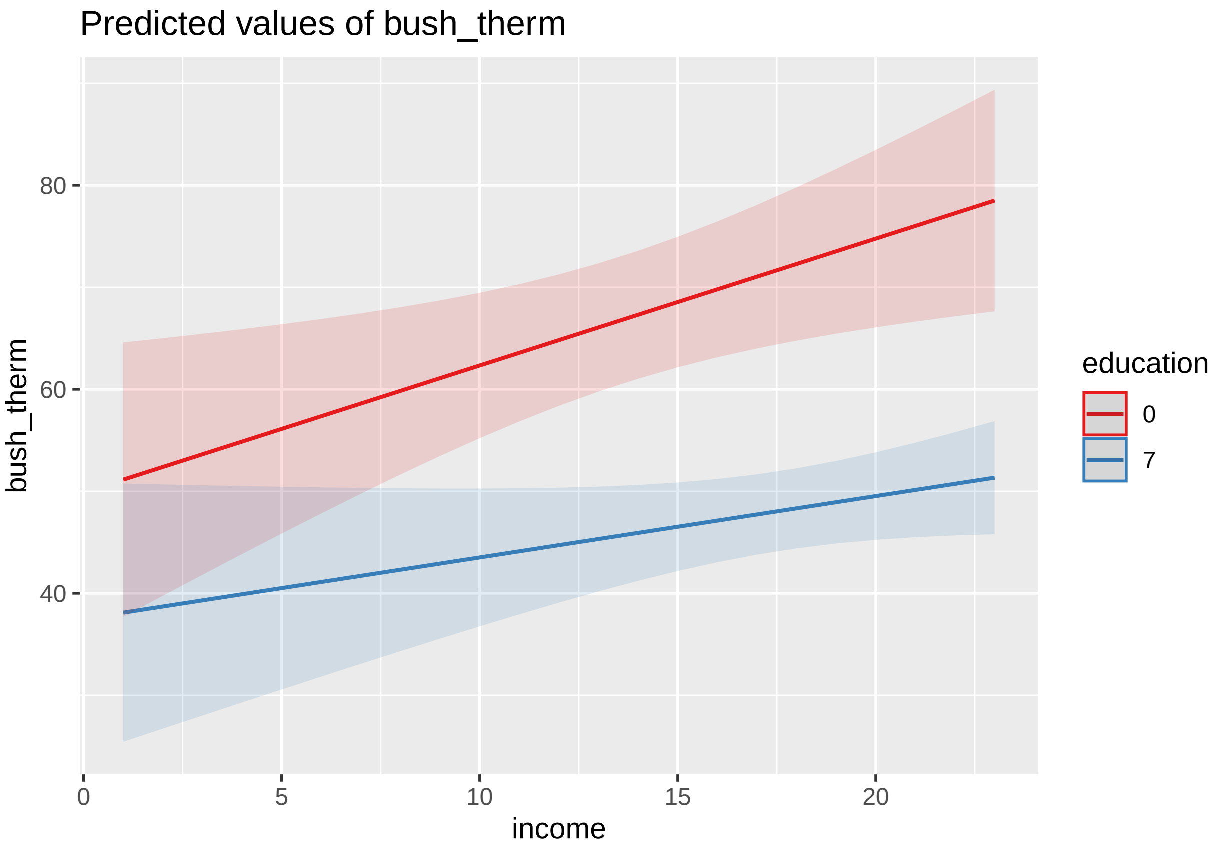
```r
#the actual plot
interplot(m=interactive.model2, var1="income", var2="education")+
  xlab("Education")+ylab("The Effect of Income on FT Score") +
  ggtitle("Marginal Effects Plot")+theme_bw()
```



As you can see here, as the respondent's education increases, the effect of the respondent's income on his or her rating of Bush on the feeling thermometer scale decreases. The important point here is that this plot shows the actual coefficient ($\hat{\beta}$) for income at every level of education. R basically calculates $\hat{\beta}$ at every possible level of education, the same way we did earlier for the `unionhouse` variable.

However, if you still do not like the two plots above, you can also use `sjPlot` to create a plot that shows the effect of income on the Bush Feeling Thermometer score at the minimum and maximum levels of education all on the same graph. You can do this by:

```
plot_model(interactive.model2, type = "int", terms=c("income", "education"))
```



This plot shows the predicted feeling thermometer score at various levels of income of the respondents when education=0 (the minimum)and when education=7 (the maximum).

In conclusion, including categorical variables and interaction terms can be very useful in terms of analysis and are relatively easy to incorporate in R. However, both of these inclusions cause issues for interpretation. We need to do more to interpret these models than we normally would for a basic linear model. The tools presented here can help you overcome these issues in interpretation.