Categorical Predictors and Interactions Application in R

```
library(car)
library(ggplot2)
library(effects)
library(mosaic)
```

Illustration

The data are from an RCT for adult inpatients recruited from an detox facility. Eligible subjects were adults, who spoke Spanish or English, reported alcohol, heroin, or cocaine as their first or second drug of choice, and either resided in proximity to the primary care clinic to which they would be referred, or were homeless. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking them to primary medical care. Subjects were interviewed at baseline during their detoxification stay, and follow-up interviews were undertaken every 6 months for 2 years so there are five measurement occasions. The data are in file called help.csv in the Data folder on Canvas. More information about the data and variables are available in the codebook, which is also in the Data folder on Canvas.

```
helpdata <- read.csv("help.csv")
head(helpdata)
```

```
##   id e2b1 g1b1 i11     pcs1     mcs1 cesd1 indtot1 drugrisk1 sexrisk1 pcrec1
## 1  1   NA    0  NA 54.22583 52.23480     7       5         0        1      1
## 2  2   NA    0   8 59.56066 41.72696    11      12         0        0      0
## 3  3   NA    0  NA 58.45777 56.77131    14      99        13        4      0
## 4  4   NA    0  NA 46.60988 14.65925    44      97         0        4      0
## 5  5   NA    0  64 31.41642 40.67421    26      55         0        4      0
## 6  6   NA    0   1 43.20495 50.05917    23      12         0        4      0
##   e2b2 g1b2 i12     pcs2     mcs2 cesd2 indtot2 drugrisk2 sexrisk2 pcrec2 e2b3
## 1   NA   NA  NA       NA       NA    NA      NA        NA       NA     NA   NA
## 2   NA   NA  NA       NA       NA    NA      NA        NA       NA     NA   NA
## 3   NA   NA  NA       NA       NA    NA      NA        NA       NA     NA   NA
## 4    1   NA  NA 57.56092 23.91316    NA      NA        NA       NA      0   NA
## 5   NA    0  NA 44.83340 42.44469    27      34         0        2      0    1
## 6   NA   NA  NA       NA       NA    NA      NA        NA       NA     NA   NA
##   g1b3 i13     pcs3     mcs3 cesd3 indtot3 drugrisk3 sexrisk3 pcrec3 e2b4 g1b4
## 1    0  NA 52.06106 56.06010     8       0         0        1      2   NA    0
## 2   NA  NA       NA       NA    NA      NA        NA       NA     NA   NA   NA
## 3   NA  NA       NA       NA    NA      NA        NA       NA     NA    2    0
## 4   NA  NA       NA       NA    NA      NA        NA       NA     NA   NA    0
## 5    0  13 25.02625 50.93477    15      37         0        4      0   NA    0
## 6   NA  NA       NA       NA    NA      NA        NA       NA     NA   NA   NA
##   i14     pcs4     mcs4 cesd4 indtot4 drugrisk4 sexrisk4 pcrec4 a15a a15b d1
## 1   8 52.27281 58.04143     5      34         0        3      2    0    0  3
## 2  NA       NA       NA    NA      NA        NA       NA     NA    2    3 22
## 3   3 66.24387 13.55065    49      94        19        4      0    0    0  0
## 4  NA 57.05444 53.45058    20       5         0        4      0    0    0  2
## 5   6 44.65918 32.72392    28      58         0        2      0   15    0 12
```

```
## 6  NA       NA       NA    NA      NA        NA       NA     NA    0    0  1
##   e2b f1a f1b f1c f1d f1e f1f f1g f1h f1i f1j f1k f1l f1m f1n f1o f1p f1q f1r
## 1  NA   3   2   3   0   2   3   3   0   2   3   3   0   1   2   2   2   2   3
## 2  NA   3   2   0   3   3   2   0   0   3   0   3   0   0   3   0   0   0   2
## 3  NA   3   2   3   0   2   2   1   3   2   3   1   0   1   3   2   0   0   3
## 4   1   0   0   1   3   2   2   1   3   0   0   1   2   2   2   0  NA   2   0
## 5   1   3   0   3   3   3   3   1   3   3   2   3   2   2   3   0   3   3   3
## 6  NA   1   0   1   3   0   0   0   3   0   1   1   3   1   0   1   3   0   0
##   f1s f1t g1b i1 i2 age treat homeless      pcs       mcs cesd indtot pss_fr
## 1   3   2   1 13 26  37     1        0 58.41369 25.111990   49     39      0
## 2   0   0   1 56 62  37     1        1 36.03694 26.670307   30     43      1
## 3   2   0   0  0  0  26     0        0 74.80633  6.762923   39     41     13
## 4   0   1   0  5  5  39     0        0 61.93168 43.967880   15     28     11
## 5   3   3   0 10 13  32     0        1 37.34558 21.675755   39     38     10
## 6   0   0   0  4  4  47     1        0 46.47521 55.508991    6     29      5
##   drugrisk sexrisk satreat drinkstatus daysdrink anysubstatus daysanysub
## 1        0       4       0           1       177            1        177
## 2        0       7       0           1         2            1          2
## 3       20       2       0           1         3            1          3
## 4        0       4       1           0       196            1        189
## 5        0       6       0           1         2            1          2
## 6        0       5       0           1        31            1         31
##   linkstatus dayslink female substance racegrp
## 1          1      225      0   cocaine   black
## 2         NA       NA      0   alcohol   white
## 3          0      365      0    heroin   black
## 4          0      343      1    heroin   white
## 5          1       57      0   cocaine   black
## 6          0      365      1   cocaine   black
```

The variable `substance` indicates their drug of choice and `female` indicates whether or not they are female with 1 indicating that they are and 0 indicating that they are not. Let's begin with the `female` variable because it has only two categories. We will predict `cesd`, a measure of depression.

Binary categorical predictor

```
m1 <- lm(cesd ~ age + female, data=helpdata)
summary(m1)
```

```
##
## Call:
## lm(formula = cesd ~ age + female, data = helpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.879  -7.882   1.117   8.402  26.399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.156e+01  2.753e+00  11.464  < 2e-16 ***
## age         9.735e-04  7.534e-02   0.013 0.989696
## female      5.289e+00  1.366e+00   3.872 0.000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 12.34 on 450 degrees of freedom
## Multiple R-squared:  0.0323, Adjusted R-squared:  0.028
## F-statistic: 7.511 on 2 and 450 DF,  p-value: 0.0006185
```

We find that, holding constant age, being female results in a 5.289 increase in the CES-D depression score and that this is statistically significant $t(450) = 3.872, p < .001$. Age is not a statistically significant predictor of the depression score, $t(450) = 0.013, p = 0.990$.

Our regression equation for predicting CESD is:

$$\widehat{CESD} = \beta_0 + \beta_1 \text{age} + \gamma \text{female} \widehat{CESD} = 31.56 + 0.001 \times \text{age} + 5.289 \times \text{female}$$

Thus, the regression equation for predicting depression score for males is:

$$\widehat{CESD} = 31.56 + 0.001 \times \text{age}$$

And the regression equation for predicting depression score for females is:

$$\widehat{CESD} = 31.56 + 0.001 \times \text{age} + 5.289 = (31.56 + 5.289) + 0.001 \times \text{age} = 36.849 + 0.001 \times \text{age}$$

Age and female account of 3% of the variance in the depression score. Although this is small, it is a statistically significant proportion of the variance, $F(2, 450) = 7.511, p < .001$.

Multi-category predictor

Next, let's consider `substance`, which has three categories. This variable is coded as a factor in R.

```
str(helpdata)
```

```
## 'data.frame':    453 obs. of  88 variables:
##  $ id          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ e2b1        : int  NA NA NA NA NA NA NA 1 1 1 ...
##  $ g1b1        : int  0 0 0 0 0 0 NA 0 0 0 ...
##  $ i11         : int  NA 8 NA NA 64 1 NA NA 12 13 ...
##  $ pcs1        : num  54.2 59.6 58.5 46.6 31.4 ...
##  $ mcs1        : num  52.2 41.7 56.8 14.7 40.7 ...
##  $ cesd1       : int  7 11 14 44 26 23 NA 18 33 37 ...
##  $ indtot1     : int  5 12 99 97 55 12 NA 82 92 104 ...
##  $ drugrisk1   : int  0 0 13 0 0 0 NA 0 8 14 ...
##  $ sexrisk1    : int  1 0 4 4 4 4 NA 4 4 5 ...
##  $ pcrec1      : int  1 0 0 0 0 0 NA 0 0 2 ...
##  $ e2b2        : int  NA NA NA 1 NA NA 1 NA NA 2 ...
##  $ g1b2        : int  NA NA NA NA 0 NA 0 NA NA 0 ...
##  $ i12         : int  NA NA NA NA NA NA 13 NA NA 22 ...
##  $ pcs2        : num  NA NA NA 57.6 44.8 ...
##  $ mcs2        : num  NA NA NA 23.9 42.4 ...
##  $ cesd2       : int  NA NA NA NA 27 NA 47 NA NA 47 ...
##  $ indtot2     : int  NA NA NA NA 34 NA 92 NA NA 100 ...
##  $ drugrisk2   : int  NA NA NA NA 0 NA 0 NA NA 19 ...
##  $ sexrisk2    : int  NA NA NA NA 2 NA 9 NA NA 0 ...
##  $ pcrec2      : int  NA NA NA 0 0 NA 2 NA NA 2 ...
##  $ e2b3        : int  NA NA NA NA 1 NA NA 2 4 NA ...
##  $ g1b3        : int  0 NA NA NA 0 NA 0 0 0 NA ...
##  $ i13         : int  NA NA NA NA 13 NA 13 12 27 NA ...
```

```
##  $ pcs3        : num  52.1 NA NA NA 25 ...
##  $ mcs3        : num  56.1 NA NA NA 50.9 ...
##  $ cesd3       : int  8 NA NA NA 15 NA 54 25 35 NA ...
##  $ indtot3     : int  0 NA NA NA 37 NA 104 92 88 NA ...
##  $ drugrisk3   : int  0 NA NA NA 0 NA 0 6 7 NA ...
##  $ sexrisk3    : int  1 NA NA NA 4 NA 7 6 8 NA ...
##  $ pcrec3      : int  2 NA NA NA 0 NA 2 0 2 NA ...
##  $ e2b4        : int  NA NA 2 NA NA NA 1 NA NA NA ...
##  $ g1b4        : int  0 NA 0 0 0 NA 0 NA 0 NA ...
##  $ i14         : int  8 NA 3 NA 6 NA NA NA 9 NA ...
##  $ pcs4        : num  52.3 NA 66.2 57.1 44.7 ...
##  $ mcs4        : num  58 NA 13.6 53.5 32.7 ...
##  $ cesd4       : int  5 NA 49 20 28 NA 52 NA 27 NA ...
##  $ indtot4     : int  34 NA 94 5 58 NA 113 NA 93 NA ...
##  $ drugrisk4   : int  0 NA 19 0 0 NA 0 NA 0 NA ...
##  $ sexrisk4    : int  3 NA 4 4 2 NA 7 NA 3 NA ...
##  $ pcrec4      : int  2 NA 0 0 0 NA 2 NA 2 NA ...
##  $ a15a        : int  0 2 0 0 15 0 0 4 1 4 ...
##  $ a15b        : int  0 3 0 0 0 0 0 1 134 20 ...
##  $ d1          : int  3 22 0 2 12 1 14 1 14 4 ...
##  $ e2b         : int  NA NA NA 1 1 NA 1 8 7 3 ...
##  $ f1a         : int  3 3 3 0 3 1 3 1 3 2 ...
##  $ f1b         : int  2 2 2 0 0 0 1 1 2 3 ...
##  $ f1c         : int  3 0 3 1 3 1 3 2 3 3 ...
##  $ f1d         : int  0 3 0 3 3 3 1 3 1 0 ...
##  $ f1e         : int  2 3 2 2 3 0 3 3 3 1 ...
##  $ f1f         : int  3 2 2 2 3 0 3 3 3 2 ...
##  $ f1g         : int  3 0 1 1 1 0 3 3 3 3 ...
##  $ f1h         : int  0 0 3 3 3 3 1 1 0 0 ...
##  $ f1i         : int  2 3 2 0 3 0 3 1 3 3 ...
##  $ f1j         : int  3 0 3 0 2 1 3 0 3 1 ...
##  $ f1k         : int  3 3 1 1 3 1 3 3 3 3 ...
##  $ f1l         : int  0 0 0 2 2 3 0 1 1 0 ...
##  $ f1m         : int  1 0 1 2 2 1 0 3 2 3 ...
##  $ f1n         : int  2 3 3 2 3 0 3 0 3 3 ...
##  $ f1o         : int  2 0 2 0 0 1 3 1 2 1 ...
##  $ f1p         : int  2 0 0 NA 3 3 1 0 0 0 ...
##  $ f1q         : int  2 0 0 2 3 0 3 2 1 0 ...
##  $ f1r         : int  3 2 3 0 3 0 3 2 3 3 ...
##  $ f1s         : int  3 0 2 0 3 0 3 0 1 0 ...
##  $ f1t         : int  2 0 0 1 3 0 3 0 2 3 ...
##  $ g1b         : int  1 1 0 0 0 0 1 1 0 0 ...
##  $ i1          : int  13 56 0 5 10 4 13 12 71 20 ...
##  $ i2          : int  26 62 0 5 13 4 20 24 129 27 ...
##  $ age         : int  37 37 26 39 32 47 49 28 50 39 ...
##  $ treat       : int  1 1 0 0 0 1 0 1 0 1 ...
##  $ homeless    : int  0 1 0 0 1 0 0 1 1 1 ...
##  $ pcs         : num  58.4 36 74.8 61.9 37.3 ...
##  $ mcs         : num  25.11 26.67 6.76 43.97 21.68 ...
##  $ cesd        : int  49 30 39 15 39 6 52 32 50 46 ...
##  $ indtot      : int  39 43 41 28 38 29 38 44 44 44 ...
##  $ pss_fr      : int  0 1 13 11 10 5 1 4 5 0 ...
##  $ drugrisk    : int  0 0 20 0 0 0 0 7 18 20 ...
##  $ sexrisk     : int  4 7 2 4 6 5 8 6 8 0 ...
```

```
##  $ satreat     : int  0 0 0 1 0 0 1 1 0 1 ...
##  $ drinkstatus : int  1 1 1 0 1 1 NA 1 1 1 ...
##  $ daysdrink   : int  177 2 3 196 2 31 NA 47 62 115 ...
##  $ anysubstatus: int  1 1 1 1 1 1 NA 1 1 1 ...
##  $ daysanysub  : int  177 2 3 189 2 31 NA 47 31 115 ...
##  $ linkstatus  : int  1 NA 0 0 1 0 0 0 0 0 ...
##  $ dayslink    : int  225 NA 365 343 57 365 334 365 365 382 ...
##  $ female      : int  0 0 0 1 0 1 1 0 1 0 ...
##  $ substance   : chr  "cocaine" "alcohol" "heroin" "heroin" ...
##  $ racegrp     : chr  "black" "white" "black" "white" ...
```

When we include it in the model, R will automatically create the dummy variables for us. We do not have to do it ourselves.

```
m2 <- lm(cesd ~ age + substance, data=helpdata)
summary(m2)
```

```
##
## Call:
## lm(formula = cesd ~ age + substance, data = helpdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.6688  -8.5775   0.6909   8.5567  30.6611
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      34.94249    3.12027  11.199  < 2e-16 ***
## age              -0.01491    0.07801  -0.191 0.848490
## substancecocaine -5.00707    1.39200  -3.597 0.000358 ***
## substanceheroin   0.42719    1.48894   0.287 0.774314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.31 on 449 degrees of freedom
## Multiple R-squared:  0.03828,    Adjusted R-squared:  0.03185
## F-statistic: 5.957 on 3 and 449 DF,  p-value: 0.0005461
```

R creates two dummy variables. If the `substance=cocaine` then $D_1 = 1$, othewise $D_1 = 0$. If `substance=heroin` then $D_2 = 1$ otherwise $D_2 = 0$. Only two dummy variables are needed because if both D_1 and D_2 are 0, then that means that `substance=alcohol` and alcohol is considered the reference group.

The regression model for predicting depression is:

$$\widehat{CESD} = \beta_0 + \beta_1 \text{age} + \gamma_1 D_1 + \gamma_2 D_2 \widehat{CESD} = 34.94 - 0.015 \times \text{age} - 5.007 \times D_1 + 0.427 \times D_2$$

Thus, the regression equation for the alcohol group is:

$$\widehat{CESD} = 34.94 - 0.015 \times \text{age}$$

The regression equation for the cocaine group is:

$$\widehat{CESD} = 34.94 - 0.015 \times \text{age} - 5.007 = 34.94 - 5.007 - 0.015 \times \text{age} = 29.933 - 0.015 \times \text{age}$$

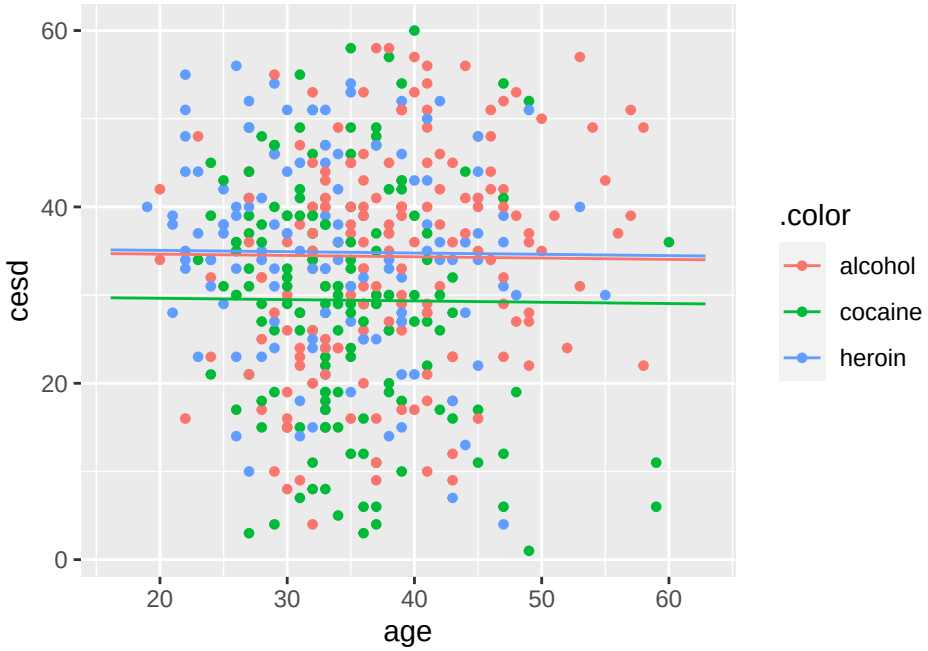
The regression equation for the heroin group is:

$$\widehat{CESD} = 34.94 - 0.015 \times \text{age} + 0.427 = 34.94 + 0.427 - 0.015 \times \text{age} = 35.367 - 0.015 \times \text{age}$$

Holding constant age, the cocaine group is statistically significantly lower on the depression scale by five points than the alcohol group, $t(449) = -3.597, p < .001$. Holding constant age, the heroin group is not statistically significantly different from the alcohol group on the depression scale, $t(449) = 0.287, p = 0.774$. There is not a significance test comparing the cocaine and heroin groups. To obtain it, you would need to change the reference substance group from alcohol to either cocaine or heroin.

Next, we can use the `plotModel` function from the `mosaic` package to plot the regression lines for each substance group. Note that this function only works if the categorical variable is coded as a factor.

```
plotModel(m2, system="ggplot2")
```



We want to test the null hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$, that there is no effect of substance on depression and we can do that using the incremental F test for a subset of slopes that we discussed previously.

```
m3 <- lm(cesd ~ age, data=helpdata)
anova(m3, m2)
```

```
## Analysis of Variance Table
##
## Model 1: cesd ~ age
## Model 2: cesd ~ age + substance
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    451 70784
## 2    449 68079  2    2704.7 8.9191 0.000159 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$, $F(2, 449) = 8.919, p < .001$. Substance has an effect on depression above and beyond age.

Interaction with binary categorical predictor

Using the HELP data, we will fit a model for predicting the CESD depression score from age, sex, and the interaction between age and sex.

$$Y_i = \alpha + \beta \times \text{age}_i + \gamma \times \text{female}_i + \delta \times (\text{age}_i \times \text{female}_i) + \epsilon_i$$

α and β are the intercept and slope for the regression of depression on age among men.

γ gives the difference in intercepts between the female and male groups

δ gives the difference in slopes between the two groups.

To test for interaction, we can test the hypothesis $H_0 : \delta = 0$

In the additive, no-interaction model, γ represented the unique partial effect of sex, while the slope β represented the unique partial effect of age.

In the interaction model, γ is no longer interpretable as the unqualified depression difference between men and women of equal age — γ is now the depression difference at $age = 0$, which in this context, does not make sense.

Likewise, in the interaction model, β is not the unqualified partial effect of age, but rather the effect of age among men.

The effect of education among women $\beta + \delta$ does not appear directly in the model.

```
m4 <- lm(cesd ~ age + female + age:female, data=helpdata)
summary(m4)
```

```
##
## Call:
## lm(formula = cesd ~ age + female + age:female, data = helpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.735  -7.784   1.199   8.415  26.411
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.726498   3.114892  10.185   <2e-16 ***
## age         -0.003615   0.085803  -0.042    0.966
## female       4.560764   6.633781   0.688    0.492
## age:female   0.020182   0.179943   0.112    0.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.35 on 449 degrees of freedom
## Multiple R-squared:  0.03233,    Adjusted R-squared:  0.02586
## F-statistic:     5 on 3 and 449 DF,  p-value: 0.002022
```

The interaction between sex and age is not statistically significant, $t(449) = 0.112, p = 0.911$. Interestingly, the effect of female is no longer significant either, $t(449) = 0.688, p = 0.492$.

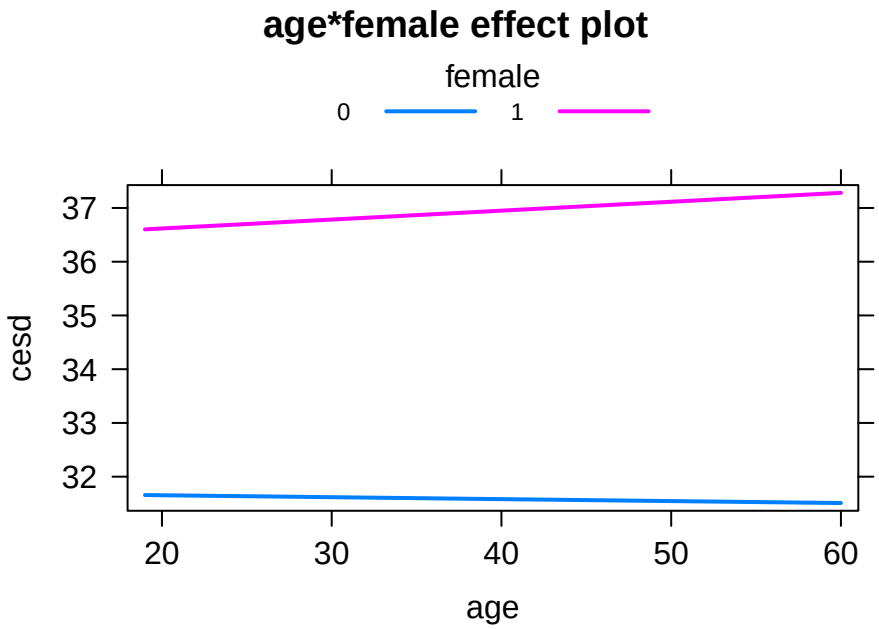
For men, female $= 0$, thus the regression equation for predicting depression is:

$$\widehat{CESD} = \alpha + \beta \text{ age} + \gamma 0 + \delta(\text{age} \times 0) = 31.726 - 0.004 \text{ age}$$

For women, female $= 1$, thus the regression equation for predicting depression is:

$$\widehat{CESD} = \alpha + \beta \text{ age} + \gamma 1 + \delta(\text{age} \times 1) = (31.726 + 4.561) + (-0.004 + 0.020) \text{ age} = 36.287 + 0.016 \text{ age}$$

```
plot(effect("age:female", m4, xlevels= list(age=19:60, female=seq(0,1))), multiline=TRUE, rug=FALSE)
```



The interaction term was automatically created by using `:` in the `lm()` function. You can shorten your code even further by using `lm(cesd ~ age*female, data=helpdata)`. If, for some reason, you would like to create and save the product term in your dataframe, you could do so one of two ways.

1. Use the interaction function

```
helpdata$intx1 <- interaction(helpdata$age, helpdata$female)
```

2. Alternatively, an interaction is the product of two variables, so just multiply them together.

```
helpdata$intx2 <- helpdata$age*helpdata$female
```

Interactions with multi-category predictor

We require one interaction term for each product of a dummy variable with a quantitative predictor variable.

If the categorical variable is a factor, then R will automatically create the dummy variables and the interactions.

```
m5 <- lm(cesd ~ age + substance + age:substance, data=helpdata)
summary(m5)
```

```
##
## Call:
## lm(formula = cesd ~ age + substance + age:substance, data = helpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.724  -8.629   0.775   8.349  32.271
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           22.1792     4.6650   4.754 2.69e-06 ***
## age                    0.3192     0.1198   2.665  0.00797 **
## substancecocaine      17.8417     6.9812   2.556  0.01093 *
## substanceheroin       19.4400     6.6356   2.930  0.00357 **
## age:substancecocaine  -0.6265     0.1903  -3.293  0.00107 **
## age:substanceheroin   -0.5210     0.1822  -2.860  0.00444 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.16 on 447 degrees of freedom
## Multiple R-squared:  0.06657,    Adjusted R-squared:  0.05613
## F-statistic: 6.376 on 5 and 447 DF,  p-value: 9.817e-06
```

The regression equation for predicting depression is:

$$\widehat{CESD} = \alpha + \beta \text{age} + \gamma_1 D_1 + \gamma_2 D_2 + \delta_1 \text{age} D_1 + \delta_2 \text{age} D_2 \widehat{CESD} = 22.179 + 0.319 \text{age} + 17.842 D_1 + 19.44 D_2 - 0.627 (\text{age}) D_1 - 0.521 (\text{a}$$

The regression equation for the alcohol group is:

$$\widehat{CESD} = 22.179 + 0.319 \text{age} + 17.842(0) + 19.44(0) - 0.627(\text{age})(0) - 0.521(\text{age})(0) = 22.179 + 0.319 \text{age}$$

For each year increase in age, the depression score increases by 0.319 points in the alcohol group and this effect is statistically significantly different from zero, $t(447) = 2.665, p = 0.008$.

The intercept, 22.179, is statistically significantly different from zero, $t(447) = 4.754, p < .001$, however, this is the predicted depression score in the alcohol group when age is zero. Thus, in this context, the intercept does not make sense because it is outside the age range in our data and hopefully newborns do not have a preferred substance that they like to use.

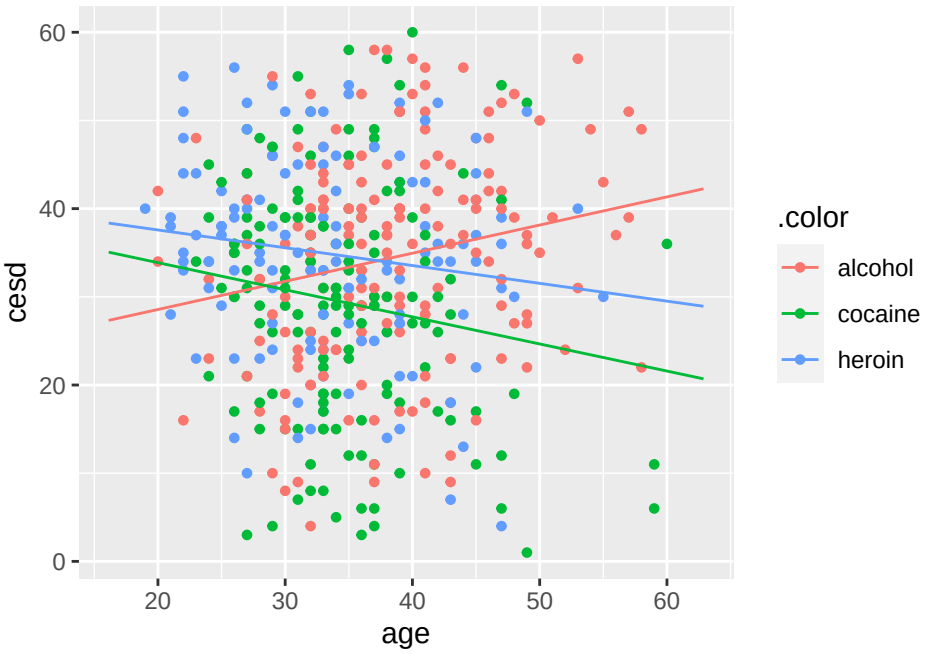
The regression equation for the cocaine group is:

$$\widehat{CESD} = 22.179 + 0.319 \text{age} + 17.842(1) + 19.44(0) - 0.627(\text{age})(1) - 0.521(\text{age})(0) = 22.179 + 17.842 + (0.319 - 0.627) \text{age} = 40.021$$

For each year increase in age, the depression score decreases by 0.308 points in the cocaine group. There is not a significance test for the null hypothesis that the slope is significantly different from zero directly in the results but we do know that the slope for the cocaine group is statistically significantly different from that of the alcohol group (from the statistical significance for -0.627, $t(447) = -3.293, p = 0.001$).

What is the regression equation for the heroin group?

```
plotModel(m5, system="ggplot2")
```



The plot illustrates that the depression score increases with increasing age in the alcohol group but the depression score decreases with increasing age in the cocaine and heroin groups and these effects are statistically different from that for the alcohol group, $t(447) = -3.293, p = 0.001$ for the cocaine group and $t(447) = -2.860, p = 0.004$ for the heroin group. We do not have a significance test for the cocaine vs. heroin groups although we could obtain one by changing the reference substance group from alcohol to either cocaine or heroin.

Hypothesis Testing

To test the null hypothesis of no interaction between age and substance, $H_0 : \delta_1 = \delta_2 = 0$, we need to delete the interaction terms from the full model and calculate an incremental F-test.

```
#m5 is the full model
#m2 is the model without the interactions
anova(m2, m5)
```

```
## Analysis of Variance Table
##
## Model 1: cesd ~ age + substance
## Model 2: cesd ~ age + substance + age:substance
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    449 68079
## 2    447 66076  2    2002.8 6.7745 0.001263 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, the interactions terms contribute significantly, $F(2, 447) = 6.775, p = .001$.

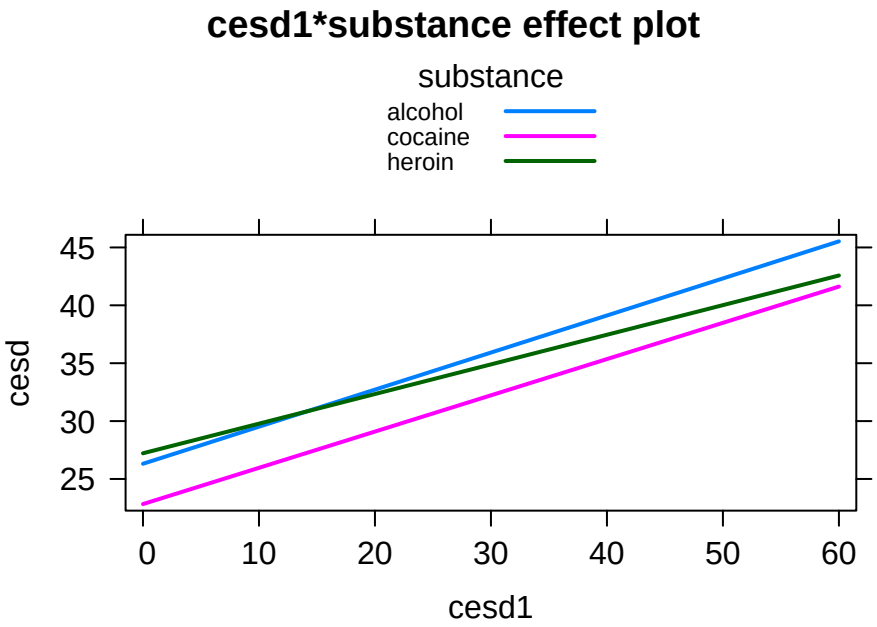
More on marginal effects

Now let's consider adding the baseline measure of depression, as we would expect baseline depression to significantly predict later depression.

```
m6 <- lm(cesd ~ age + cesd1 + substance + age:substance + cesd1:substance,
         data=helpdata)
summary(m6)
```

```
##
## Call:
## lm(formula = cesd ~ age + cesd1 + substance + age:substance +
##     cesd1:substance, data = helpdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.5144  -7.4332   0.4726   7.1895  27.4466
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              19.636420   5.632873   3.486 0.000584 ***
## age                       0.183886   0.139240   1.321 0.187894
## cesd1                     0.320193   0.078522   4.078 6.21e-05 ***
## substancecocaine         13.460401   9.287915   1.449 0.148593
## substanceheroin          16.651799   8.511516   1.956 0.051594 .
## age:substancecocaine     -0.466626   0.234283  -1.992 0.047550 *
## age:substanceheroin      -0.433663   0.222065  -1.953 0.052013 .
## cesd1:substancecocaine   -0.007046   0.120496  -0.058 0.953418
## cesd1:substanceheroin    -0.064253   0.123126  -0.522 0.602263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.06 on 237 degrees of freedom
##   (207 observations deleted due to missingness)
## Multiple R-squared:  0.1935, Adjusted R-squared:  0.1663
## F-statistic: 7.108 on 8 and 237 DF,  p-value: 2.009e-08
```

```
plot(effect("cesd1:substance", m6, xlevels=list(cesd1=0:60)), multiline=TRUE, rug=FALSE)
```



The regression model for predicting depression is:

$$\widehat{CESD} = \alpha + \beta_1 \text{age} + \beta_2 \text{cesd} + \gamma_1 D_1 + \gamma_2 D_2 + \delta_{11} \text{age} D_1 + \delta_{12} \text{age} D_2 + \delta_{21} \text{cesd} D_1 + \delta_{22} \text{cesd} D_2$$

To test the null hypothesis of no interaction between baseline depression and substance, $H_0 : \delta_{21} = \delta_{22} = 0$,

we delete the interaction terms from the full model.

```
m7 <- lm(cesd ~ age + cesd1 + substance + age:substance, data=helpdata)
summary(m7)
```

```
##
## Call:
## lm(formula = cesd ~ age + cesd1 + substance + age:substance,
##     data = helpdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.2393  -7.5436   0.6057   7.3080  27.1324
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           19.90749    5.55553   3.583 0.000411 ***
## age                    0.18903    0.13791   1.371 0.171769
## cesd1                  0.29987    0.05026   5.967 8.68e-09 ***
## substancecocaine      13.69417    8.60845   1.591 0.112980
## substanceheroin       15.34893    8.13487   1.887 0.060398 .
## age:substancecocaine  -0.47831    0.23130  -2.068 0.039722 *
## age:substanceheroin   -0.44044    0.22063  -1.996 0.047032 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.02 on 239 degrees of freedom
##   (207 observations deleted due to missingness)
## Multiple R-squared:  0.1925, Adjusted R-squared:  0.1722
## F-statistic: 9.495 on 6 and 239 DF,  p-value: 2.335e-09
```

```
anova(m7, m6)
```

```
## Analysis of Variance Table
##
## Model 1: cesd ~ age + cesd1 + substance + age:substance
## Model 2: cesd ~ age + cesd1 + substance + age:substance + cesd1:substance
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    239 29024
## 2    237 28987  2    37.016 0.1513 0.8597
```

We fail to reject the null hypothesis, $F(2, 237) = 0.1513, p = 0.8597$. It does not appear that the interaction between baseline depression and substance contributes above and beyond the other variables in the model.

The logic of the principle of marginality is to interpret the interaction first. Conforming to the principle of marginality, the test for each main effect is computed assuming that the interactions that are higher-order relatives of that main effect are 0.

Thus, for example, the test for the age main effect assumes that the age-by-substance interaction is absent (i.e., that $\delta_{11} = \delta_{12} = 0$), but the test for the main effect of baseline depression does not assume that the age-by-substance interaction is absent.

Tests formulated according to the principle of marginality are sometimes called Type II tests.

The degrees of freedom for the several sources of variation add to the total degrees of freedom, but — because the regressors in different sets are correlated — the sums of squares do not add to the total sum of squares. What is important is that sensible hypotheses are tested, not that the sums of squares add to the total sum of squares.

The principle of marginality serves as a guide to constructing incremental F-tests for the terms in a model that includes interactions.

Summary

A dichotomous or binary categorical predictor variable can be entered into a regression equation by formulating a dummy variable, coded 1 for one category of the variable and 0 for the other category.

A multi-category predictor variable can be entered into a regression by coding a set of 0/1 dummy variables, one fewer than the number of categories of the variable.

The 'omitted' category, coded 0 for all dummy variables in the set, serves as the reference group.

Interactions can be incorporated by computing interaction terms by taking products of dummy variables with quantitative predictor variables.

The model permits different slopes for different subgroups — that is, regression surfaces that are not parallel.

The principle of marginality specifies that a model including a higher order term (such as an interaction) should normally also include the lower-order relatives of that term (the main effects that 'compose' the interaction).