26.3: Interactions Between Variables

In the previous model, we assumed that the effect of study time on grade (i.e., the regression slope) was the same for both groups. However, in some cases we might imagine that the effect of one variable might differ depending on the value of another variable, which we refer to as an *interaction* between variables.

Let's use a new example that asks the question: What is the effect of caffeine on public speaking? First let's generate some data and plot them. Looking at panel A of Figure 26.4, there doesn't seem to be a relationship, and we can confirm that by performing linear regression on the data:

```
##
## Call:
## lm(formula = speaking ~ caffeine, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33.10 -16.02   5.01  16.45  26.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.413      9.165   -0.81     0.43
## caffeine       0.168      0.151    1.11     0.28
##
## Residual standard error: 19 on 18 degrees of freedom
## Multiple R-squared:  0.0642, Adjusted R-squared:  0.0122
## F-statistic: 1.23 on 1 and 18 DF,  p-value: 0.281
```

But now let's say that we find research suggesting that anxious and non-anxious people react differently to caffeine. First let's plot the data separately for anxious and non-anxious people.

As we see from panel B in Figure 26.4, it appears that the relationship between speaking and caffeine is different for the two groups, with caffeine improving performance for people without anxiety and degrading performance for those with anxiety. We'd like to create a statistical model that addresses this question. First let's see what happens if we just include anxiety in the model.

```
##
## Call:
## lm(formula = speaking ~ caffeine + anxiety, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.97  -9.74   1.35  10.53  25.36
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -12.581      9.197   -1.37     0.19
## caffeine             0.131      0.145    0.91     0.38
## anxietynotAnxious   14.233      8.232    1.73     0.10
##
## Residual standard error: 18 on 17 degrees of freedom
## Multiple R-squared:  0.204,  Adjusted R-squared:  0.11
## F-statistic: 2.18 on 2 and 17 DF,  p-value: 0.144
```

Here we see there are no significant effects of either caffeine or anxiety, which might seem a bit confusing. The problem is that this model is trying to fit the same line relating speaking to caffeine for both groups. If we want to fit them using separate lines, we need to include an *interaction* in the model, which is equivalent to fitting different lines for each of the two groups; in R this is denoted by the * symbol.

```
##
## Call:
## lm(formula = speaking ~ caffeine + anxiety + caffeine * anxiety,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.385  -7.103  -0.444   6.171  13.458
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 17.4308     5.4301    3.21  0.00546 **
## caffeine                    -0.4742     0.0966   -4.91  0.00016 ***
## anxietynotAnxious          -43.4487     7.7914   -5.58  4.2e-05 ***
## caffeine:anxietynotAnxious   1.0839     0.1293    8.38  3.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.1 on 16 degrees of freedom
## Multiple R-squared:  0.852,  Adjusted R-squared:  0.825
## F-statistic: 30.8 on 3 and 16 DF,  p-value: 7.01e-07
```

From these results we see that there are significant effects of both caffeine and anxiety (which we call *main effects*) and an interaction between caffeine and anxiety. Panel C in Figure 26.4 shows the separate regression lines for each group.

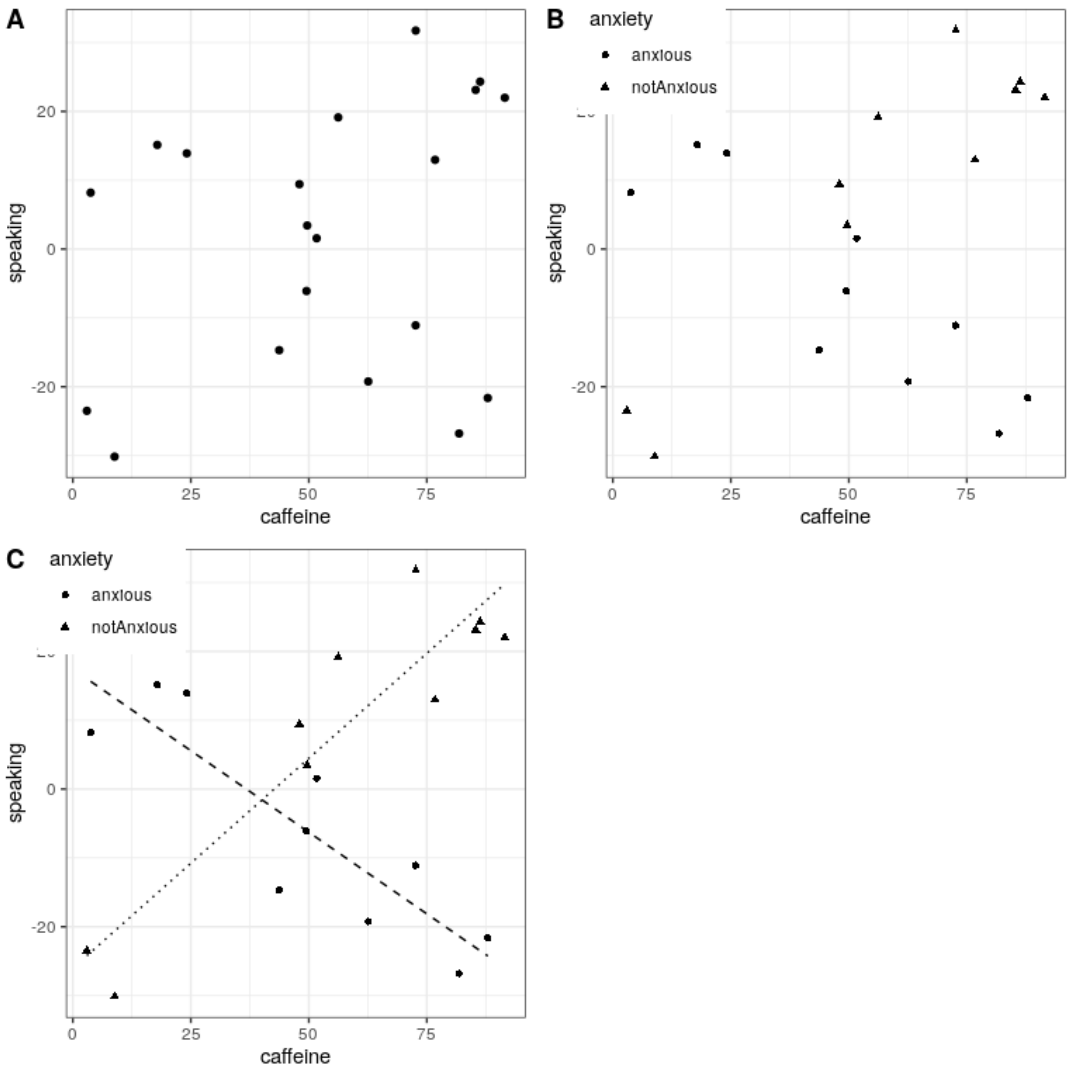

Figure 26.4: A: The relationship between caffeine and public speaking. B: The relationship between caffeine and public speaking, with anxiety represented by the shape of the data points. C: The relationship between public speaking and caffeine, including an interaction with anxiety. This results in two lines that separately model the slope for each group (dashed for anxious, dotted for non-anxious).

Sometimes we want to compare the relative fit of two different models, in order to determine which is a better model; we refer to this as *model comparison*. For the models above, we can compare the goodness of fit of the model with and without the interaction, using the `anova()` command in R:

```
## Analysis of Variance Table
##
## Model 1: speaking ~ caffeine + anxiety
## Model 2: speaking ~ caffeine + anxiety + caffeine * anxiety
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     17 5639
## 2     16 1046  1      4593 70.3  3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This tells us that there is good evidence to prefer the model with the interaction over the one without an interaction. Model comparison is relatively simple in this case because the two models are *nested* – one of the models is a simplified version of the other model. Model comparison with non-nested models can get much more complicated.