Moderation

Utrecht University Winter School: Regression in R

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Moderation

So far we've been discussing *additive models*.

- Additive models allow us to examine the partial effects of several predictors on some outcome.
  - The effect of one predictor does not change based on the values of other predictors.

Now, we'll discuss *moderation*.

- Moderation allows us to ask *when* one variable, X, affects another variable, Y.
  - We're considering the conditional effects of X on Y given certain levels of a third variable Z.

Equations

In additive MLR, we might have the following equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

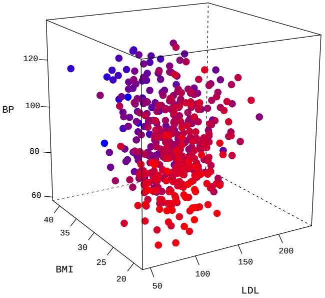
This additive equation assumes that X and Z are independent predictors of Y.

When X and Z are independent predictors, the following are true:

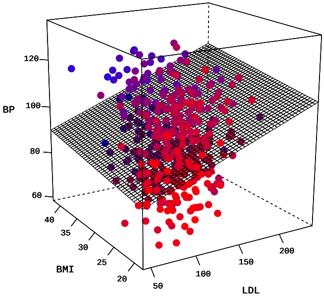
- X and Z *can* be correlated.
- β_1 and β_2 are *partial* regression coefficients.
- The effect of X on Y is the same at **all levels** of Z, and the effect of Z on Y is the same at **all levels** of X.

Additive Regression

The effect of X on Y is the same at **all levels** of Z.

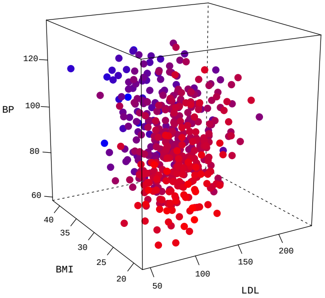


$\longrightarrow$

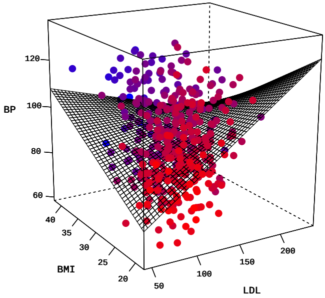


Moderated Regression

The effect of X on Y varies **as a function** of Z.



$\longrightarrow$



Equations

The following derivation is adapted from Hayes (2018).

- When testing moderation, we hypothesize that the effect of X on Y varies as a function of Z.
- We can represent this concept with the following equation:

$$Y = \beta_0 + f(Z)X + \beta_2 Z + \varepsilon \tag{1}$$

Equations

The following derivation is adapted from Hayes (2018).

- When testing moderation, we hypothesize that the effect of X on Y varies as a function of Z.
- We can represent this concept with the following equation:

$$Y = \beta_0 + f(Z)X + \beta_2 Z + \varepsilon \tag{1}$$

- If we assume that Z linearly (and deterministically) affects the relationship between X and Y, then we can take:

$$f(Z) = \beta_1 + \beta_3 Z \tag{2}$$

Equations

- Substituting Equation 2 into Equation 1 leads to:

$$Y = \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon$$

Equations

- Substituting Equation 2 into Equation 1 leads to:

$$Y = \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon$$

- Which, after distributing X and reordering terms, becomes:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

Interpretation

Given the following equation:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z + \hat{\beta}_3 XZ + \hat{\varepsilon}$$

- $\hat{\beta}_3$ quantifies the effect of Z on the focal effect (the $X \rightarrow Y$ effect).
  - For a unit change in Z, $\hat{\beta}_3$ is the expected change in the effect of X on Y.

- $\hat{\beta}_1$ and $\hat{\beta}_2$ are *conditional effects*.
  - Interpreted where the other predictor is zero.
  - For a unit change in X, $\hat{\beta}_1$ is the expected change in Y, when $Z = 0$.
  - For a unit change in Z, $\hat{\beta}_2$ is the expected change in Y, when $X = 0$.

Example

Still looking at the *diabetes* dataset.

- We suspect that patients' BMIs are predictive of their average blood pressure.
- We further suspect that this effect may be differentially expressed depending on the patients' LDL levels.

Example

```
## Focal Effect:
out0 <- lm(bp ~ bmi, data = dDat)
partSummary(out0, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.9973     3.6659   16.91   <2e-16
bmi           1.2379     0.1371    9.03   <2e-16

Residual standard error: 12.72 on 440 degrees of freedom
Multiple R-squared:  0.1563, Adjusted R-squared:  0.1544
F-statistic: 81.54 on 1 and 440 DF,  p-value: < 2.2e-16
```

Example

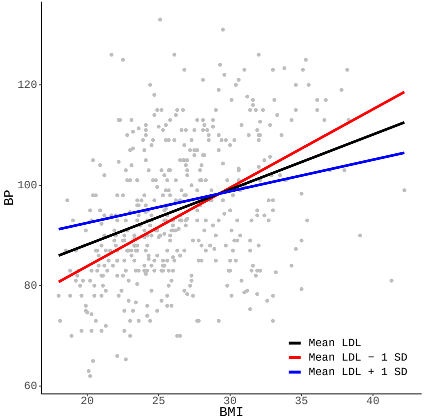
```
## Additive Model:
out1 <- lm(bp ~ bmi + ldl, data = dDat)
partSummary(out1, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.26577    3.91281  15.147  < 2e-16
bmi          1.16567    0.14156   8.235 2.08e-15
ldl          0.04016    0.02056   1.953   0.0515

Residual standard error: 12.68 on 439 degrees of freedom
Multiple R-squared:  0.1636, Adjusted R-squared:  0.1598
F-statistic: 42.94 on 2 and 439 DF,  p-value: < 2.2e-16
```

Example

```
## Moderated Model:
out2 <- lm(bp ~ bmi * ldl, data = dDat)
partSummary(out2, -c(1, 2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.480616  14.291677   1.013 0.311514
bmi          2.867825   0.541312   5.298 1.86e-07
ldl          0.448771   0.127160   3.529 0.000461
bmi:ldl     -0.015352   0.004716  -3.255 0.001221

Residual standard error: 12.54 on 438 degrees of freedom
Multiple R-squared:  0.1834, Adjusted R-squared:  0.1778
F-statistic: 32.78 on 3 and 438 DF,  p-value: < 2.2e-16
```

Visualizing the Interaction

We can get a better idea of the patterns of moderation by plotting the focal effect at conditional values of the moderator.



Probing the Interaction

A significant estimate of β_3 tells us that the effect of X on Y depends on the level of Z, but nothing more.

- The plot on the previous slide gives a descriptive illustration of the pattern, but does not support statistical inference.
  - The three conditional effects we plotted look different, but we cannot say much about how they differ with only the plot and $\hat{\beta}_3$.
- This is the purpose of *probing* the interaction.
  - Try to isolate areas of Z's distribution in which $X \rightarrow Y$ effect is significant and areas where it is not.

Probing the Interaction

The most popular method of probing interactions is to do a so-called *simple slopes* analysis.

- Pick-a-point approach
- Spotlight analysis

In simple slopes analysis, we test if the slopes of the conditional effects plotted above are significantly different from zero.

- To do so, we test the significance of *simple slopes*.

Simple Slopes

Recall the derivation of our moderated equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

We can reverse the process by factoring out X and reordering terms:

$$Y = \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon$$

Where $f(Z) = \beta_1 + \beta_3 Z$ is the linear function that shows how the relationship between X and Y changes as a function of Z.

$f(Z)$ is the *simple slope*.

- By plugging different values of Z into $f(Z)$, we get the value of the conditional effect of X on Y at the chosen level of Z.

Significance Testing of Simple Slopes

The values of Z used to define the simple slopes are arbitrary.

- The most common choice is: $\left\{(\bar{Z} - SD_Z), \bar{Z}, (\bar{Z} + SD_Z)\right\}$
- You could also use interesting percentiles of Z's distribution.

The standard error of a simple slope is given by:

$$SE_{f(Z)} = \sqrt{SE^2_{\beta_1} + 2Z \cdot \text{COV}(\beta_1, \beta_3) + Z^2 SE^2_{\beta_3}}$$

So, you can test the significance of a simple slope by constructing a Wald statistic or confidence interval using $\hat{f}(Z)$ and $SE_{f(Z)}$:

$$t = \frac{\hat{f}(Z)}{SE_{f(Z)}}, \quad CI = \hat{f}(Z) \pm t_{crit} \times SE_{f(Z)}$$

Interaction Probing

When probed the interaction with simple slopes analysis:

1. Choose interesting values of the moderator, Z.
2. Check the significance of the focal effect, $X \rightarrow Y$, at the Z values chosen in Step 1.
3. Use the results from Step 2 to get an idea of where in Z's distribution the focal effect is or is not significant.

Interaction Probing

When probed the interaction with simple slopes analysis:

1. Choose interesting values of the moderator, Z.
2. Check the significance of the focal effect, $X \rightarrow Y$, at the Z values chosen in Step 1.
3. Use the results from Step 2 to get an idea of where in Z's distribution the focal effect is or is not significant.

We saw manual calculations for the the quantities needed, but there is a simpler way:

- Centering

Centering

Centering shifts the scale of a variable up or down by subtracting a constant (e.g., the variable's mean) from each of its observations.

- The most familiar form of center is *mean centering*.
- We can center on any value.
  - When probing interactions, we can center Z on the interesting values we choose to define the simple slopes.
  - Due to the interpretation of conditional effects, running the model with Z centered on a specific value automatically provides a test of the simple slope for that value of Z.

Probing via Centering

Say we want to do a simple slopes analysis to test the conditional effect of X on Y at three levels of $Z = \{Z_1, Z_2, Z_3\}$.

- All we need to do is fit the following three models:

$$Y = \beta_0 + \beta_1 X + \beta_2 (Z - Z_1) + \beta_3 X (Z - Z_1) + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 (Z - Z_2) + \beta_3 X (Z - Z_2) + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 (Z - Z_3) + \beta_3 X (Z - Z_3) + \varepsilon$$

Probing via Centering

Say we want to do a simple slopes analysis to test the conditional effect of X on Y at three levels of $Z = \{Z_1, Z_2, Z_3\}$.

- All we need to do is fit the following three models:

$$Y = \beta_0 + \beta_1 X + \beta_2 (Z - Z_1) + \beta_3 X (Z - Z_1) + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 (Z - Z_2) + \beta_3 X (Z - Z_2) + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 (Z - Z_3) + \beta_3 X (Z - Z_3) + \varepsilon$$

- The default output for $\hat{\beta}_1$ provides tests of the simple slopes.

Example

Create transformed predictors by centering on critical values of the moderator, Z_{LDL}.

```
zMean      <- mean(dDat$ldl)
zSD        <- sd(dDat$ldl)
dDat$zCen <- dDat$ldl - zMean
dDat$zHi  <- dDat$ldl - (zMean + zSD)
dDat$zLo  <- dDat$ldl - (zMean - zSD)
```

Example

Test the simple slope of $X_{BMI} \rightarrow Y_{BP}$ at 1 SD below the mean of Z_{LDL}.

```
out2.1 <- lm(bp ~ bmi*zLo, data = dDat)
partSummary(out2.1, -c(1, 2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.637886   4.764883  11.047  < 2e-16
bmi          1.562525   0.185667   8.416 5.59e-16
zLo          0.448771   0.127160   3.529 0.000461
bmi:zLo     -0.015352   0.004716  -3.255 0.001221

Residual standard error: 12.54 on 438 degrees of freedom
Multiple R-squared:  0.1834, Adjusted R-squared:  0.1778
F-statistic: 32.78 on 3 and 438 DF,  p-value: < 2.2e-16
```

The estimated slope for `bmi`, $\hat{\beta}_1 = 1.563$, is the simple slope.

Example

Test the simple slope of $X_{BMI} \rightarrow Y_{BP}$ at the mean of Z_{LDL}.

```
out2.2 <- lm(bp ~ bmi*zCen, data = dDat)
partSummary(out2.2, -c(1, 2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.286409   3.812480  17.387  < 2e-16
bmi          1.095631   0.141679   7.733 7.27e-14
zCen         0.448771   0.127160   3.529 0.000461
bmi:zCen    -0.015352   0.004716  -3.255 0.001221

Residual standard error: 12.54 on 438 degrees of freedom
Multiple R-squared:  0.1834, Adjusted R-squared:  0.1778
F-statistic: 32.78 on 3 and 438 DF,  p-value: < 2.2e-16
```

The estimated slope for `bmi`, $\hat{\beta}_1 = 1.096$, is the simple slope.

Example

Test the simple slope of $X_{BMI} \rightarrow Y_{BP}$ at 1 SD above the mean of Z_{LDL}.

```
out2.3 <- lm(bp ~ bmi*zHi, data = dDat)
partSummary(out2.3, -c(1, 2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 79.934933   6.023132  13.271  < 2e-16
bmi          0.628736   0.216369   2.906 0.003848
zHi          0.448771   0.127160   3.529 0.000461
bmi:zHi     -0.015352   0.004716  -3.255 0.001221

Residual standard error: 12.54 on 438 degrees of freedom
Multiple R-squared:  0.1834, Adjusted R-squared:  0.1778
F-statistic: 32.78 on 3 and 438 DF,  p-value: < 2.2e-16
```

The estimated slope for `bmi`, $\hat{\beta}_1 = 0.629$, is the simple slope.

Compare Approaches

The manual and the centering approaches give identical answers, barring rounding errors:

	Z Low	Z Center	Z High
Manual	1.562525	1.095631	0.628736
Centering	1.562525	1.095631	0.628736

Simple Slopes

	Z Low	Z Center	Z High
Manual	0.185667	0.141679	0.216369
Centering	0.185667	0.141679	0.216369

Standard Errors

Categorical Moderators

Categorical moderators encode *group-specific* effects.

- E.g., if we include *sex* as a moderator, we are modeling separate focal effects for males and females.

Given a set of codes representing our moderator, we specify the interactions as before:

$$Y_{total} = \beta_0 + \beta_1 X_{inten} + \beta_2 Z_{male} + \beta_3 X_{inten} Z_{male} + \varepsilon$$

$$\begin{aligned} Y_{total} = \beta_0 &+ \beta_1 X_{inten} + \beta_2 Z_{lo} + \beta_3 Z_{mid} + \beta_4 Z_{hi} \\ &+ \beta_5 X_{inten} Z_{lo} + \beta_6 X_{inten} Z_{mid} + \beta_7 X_{inten} Z_{hi} + \varepsilon \end{aligned}$$

Example

```
## Load data:
socSup <- readRDS(paste0(dataDir, "social_support.rds"))

## Focal effect:
out3 <- lm(bdi ~ tanSat, data = socSup)
partSummary(out3, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.4089     5.3502   4.562 1.54e-05
tanSat       -0.8100     0.3124  -2.593   0.0111

Residual standard error: 9.278 on 93 degrees of freedom
Multiple R-squared:  0.06742, Adjusted R-squared:  0.05739
F-statistic: 6.723 on 1 and 93 DF,  p-value: 0.01105
```

Example

```r
## Estimate the interaction:
out4 <- lm(bdi ~ tanSat * sex, data = socSup)
partSummary(out4, -c(1, 2))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     20.8478     6.2114   3.356  0.00115
tanSat          -0.5772     0.3614  -1.597  0.11372
sexmale         14.3667    12.2054   1.177  0.24223
tanSat:sexmale  -0.9482     0.7177  -1.321  0.18978

Residual standard error: 9.267 on 91 degrees of freedom
Multiple R-squared:  0.08955, Adjusted R-squared:  0.05954
F-statistic: 2.984 on 3 and 91 DF,  p-value: 0.03537
```

Example

On the last slide, the estimated slope for `tanSat`, $\hat{\beta}_1 = -0.577$, is the simple slope for females.

- To estimate the simple slope for males, we simply change the reference group of the `sex` factor and re-estimate the model.

```r
## Test the 'male' simple slope by changing reference group:
socSup$sex2 <- relevel(socSup$sex, ref = "male")

## Re-estimate the interaction:
out5 <- lm(bdi ~ tanSat * sex2, data = socSup)
```

Example

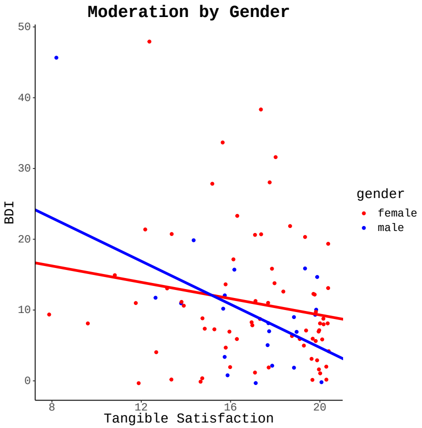
```
partSummary(out5, -c(1, 2))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         35.2146    10.5067   3.352  0.00117
tanSat              -1.5254     0.6201  -2.460  0.01579
sex2female         -14.3667    12.2054  -1.177  0.24223
tanSat:sex2female    0.9482     0.7177   1.321  0.18978

Residual standard error: 9.267 on 91 degrees of freedom
Multiple R-squared:  0.08955, Adjusted R-squared:  0.05954
F-statistic: 2.984 on 3 and 91 DF,  p-value: 0.03537
```

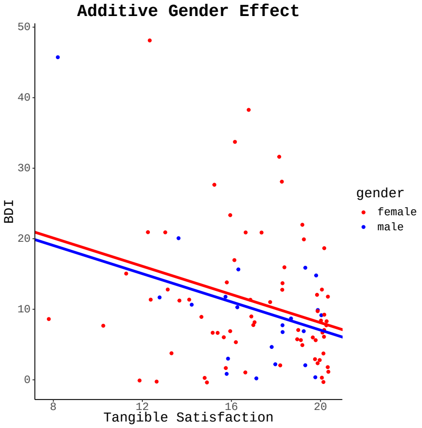
The estimated slope for `tanSat`, $\hat{\beta}_1 = -1.525$, is now the simple slope for males.

Visualizing Categorical Moderation

$$\hat{Y}_{BDI} = 20.85 - 0.58X_{tsat} + 14.37Z_{male} - 0.95X_{tsat}Z_{male}$$

$$\hat{Y}_{BDI} = 28.10 - 1.00X_{tsat} - 1.05Z_{male}$$





References

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York: Guilford Press.