Dummy Variables and Multiplicative Regression

Brad Jones

\(^1\)Department of Political Science
University of California, Davis

February 24, 2010
Today: Dummy variables and multiplicative regression
Dummy Variables

- Begin with garden variety regression:

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}, \]

- If \( X_1 \) and \( X_2 \) are continuous, then the natural interpretation for a slope coefficient is forthcoming: for an incremental (or unit) change in \( X \), the expected value of \( Y \) changes (increases or decreases) by about \( \hat{\beta} \) amount.

- What about “qualitative” or nominal outcomes? Such a variable may represent gender, racial classifications, party classifications, or so on. Fortunately, such covariates pose few problems for the regression model. The main issues involve coding categorical variables and interpreting them.
Consider only dichotomous categorical variables:

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_{1i}, \]

the predicted value of \( \hat{Y} \) when \( D = 1 \) is given by

\[ \hat{\beta}_0 + \hat{\beta}_1 1. \]

When \( D = 0 \), the predicted value of \( Y \) is given by

\[ \hat{\beta}_0 + \hat{\beta}_1 0 \]

or \( \hat{\beta}_0 \).

In a bivariate model with one dummy variable, there will only be two predicted values.

Note also that since our dummy variable has a natural 0 point (which we arbitrarily defined), the constant term as a natural interpretation: it gives us the predicted \( Y \) for the condition when \( D = 0 \).
Dummy Variables

- Dummy variables, constructed in this way, give nothing more than differences in intercepts.
- In the previous model, the coefficient $\hat{\beta}_1$ tells us the expected increase (if the coefficient is positive) in $Y$ for observations where $D = 1$.
- Note also that since the model above has no continuous (or quantitative) covariates, there are no slope coefficients. Thus, if you plotted $\hat{Y}$ with respect to $D$, you would simply get “two dots.”
- Sometimes, we include categorical covariates to account for variation in $Y$ that is a function of some group-specific or attribute-specific factor.
Dummy Variables

Consider a model with one quantitative variable and one dummy variable:

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 D_{1i}, \]

where \( \hat{\beta}_1 \) is the coefficient for our quantitative variable \( X \) and \( \hat{\beta}_2 \) is the coefficient for our dummy variable \( D \).

Unlike the previous model which yielded only differences in intercepts, this model will give us different intercepts but parallel slopes.

Suppose that \( D = 1 \), then the regression response function for this condition is:

\[ \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 1. \]
Dummy Variables

- Rearranging, we find the following:
  \[ (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X_{1i}. \]

In contrast, when \( D = 0 \), the regression response function is given by

\[ \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 0, \]

which is equivalent to

\[ (\hat{\beta}_0) + \hat{\beta}_1 X_{1i}. \]

- It is clear, the “effect” of the dummy variable in the model is to act as a “contrast” or “offset” for the two groups.
- When \( D = 1 \), the coefficient \( \hat{\beta}_2 \) represents the increase in the \( y \)-intercept attributable to \( D \); when \( D = 0 \), the constant term (i.e. \( \hat{\beta}_0 \)) gives us the intercept for this condition.
- The important thing to note, however, is the following: the effect of \( X_1 \) on \( Y \) is the same for both groups.
Dummy Variables

- The dummy variable only reflects a difference in intercepts while the slope coefficient will be equivalent for the two groups.
- Example using some simulated data.
- Fitted model:
  \[ \hat{y} = 78.9 = .95x \]
- Fitted model and fitted values plot (next two slides)
## Dummy Variables

|        | Estimate | Std. Error | t value | Pr(>|t|) |
|--------|----------|------------|---------|----------|
| (Intercept) | 78.8394 | 1.6539     | 47.67   | 0.0000   |
| x1      | 0.9520   | 0.0518     | 18.39   | 0.0000   |
Dummy Variables

Regression of Y on X1

Fitted Values

0 10 20 30 40 50 60
0 80 90 100 110 120 130
Dummy Variables

- Model has standard interpretation as with any OLS model.
- No attempt is made to account for group differences.
- One way to approach this might be for separate models by group (model for $D = 1$ and for $D = 0$)
- Model 1 then Model 2:
## Dummy Variables

|          | Estimate | Std. Error | t value | Pr(>|t|)  |
|----------|----------|------------|---------|-----------|
| (Intercept) | 79.0000  | 5.6657     | 13.94   | 0.0008    |
| x1       | 0.9500   | 0.1335     | 7.11    | 0.0057    |

|          | Estimate | Std. Error | t value | Pr(>|t|)  |
|----------|----------|------------|---------|-----------|
| (Intercept) | 78.9659  | 1.8341     | 43.05   | 0.0000    |
| x1       | 0.9340   | 0.1178     | 7.93    | 0.0042    |
Dummy Variables

Regression of Y on X1

[Graph showing a linear relationship between X1 and fitted values]
Dummy Variables

- Apparently no group-based differences.
- Separate models yield relatively similar slopes . . .
- though we still have two models.
- Incorporate $D$ into the model returns:
## Dummy Variables

|          | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 78.8451  | 1.7670     | 44.62   | 0.0000   |
| x1       | 0.9444   | 0.0840     | 11.24   | 0.0000   |
| d        | 0.3794   | 3.1682     | 0.12    | 0.9080   |
Dummy Variables

Regression of Y on X1 and D
Dummy Variables

Regression of $Y$ on $X_1$ and $D$
Dummy Variables

- The single model returns fitted values essentially identical to the separate models approach.
- There are no apparent differences due to the “groups.”
- “Controlling” for the group is superfluous: $x_1$ gives similar slopes for the two groups.
- Suppose we went the other way and ignored $x_1$?
- That is: $y = \beta_0 + \beta_1 D$
Dummy Variables

|               | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 89.8000  | 6.0183     | 14.92   | 0.0000   |
| d             | 27.2000  | 8.5112     | 3.20    | 0.0127   |
Dummy Variables

- $d$ is significantly different from 0 which suggests substantial group differences.
- $d$ is picking up on the fact that the average value of $Y \mid D$ is higher for one group versus the other.
- However, the linear association between $Y$ and $X_1$ (which is not included) is virtually identical across the two groups.
- Once we account for the variable $X$, the factor $D$ has no impact.*
- Another example.
- Consider this plot:
Dummy Variables

Scatterplot of Y and X2
Dummy Variables

- Two sets of observations both having a positive association between $Y$ and $X_2$.
- Correlation for “group 1”: .97; for “group 2”: .99
- Individually, knowing $X_2$ for each group almost perfectly predicts the outcome $Y$.
- Try the separate regressions approach again $(y = f(X_2 \mid D = 1)$ and $y = f(X_2 \mid D = 0)$
Dummy Variables

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 79.0000  | 5.6657     | 13.94   | 0.0008   |
| x2             | 0.9500   | 0.1335     | 7.11    | 0.0057   |

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 12.8000  | 3.3882     | 3.78    | 0.0325   |
| x2             | 0.7000   | 0.0306     | 22.91   | 0.0002   |
Dummy Variables

Regression of Y on X2
Dummy Variables

- Separate regressions obviously confirm the correlational analysis.
- Since we know more data are better than less data, suppose we pool the groups (like we did in the previous example).
- That is: regress $y$ on $x_2$
- The model:
## Dummy Variables

|       | Estimate | Std. Error | t value | Pr(>|t|) |
|-------|----------|------------|---------|----------|
| (Intercept) | 119.7684 | 12.7281    | 9.41    | 0.0000   |
| x2     | −0.2182  | 0.1516     | −1.44   | 0.1879   |
A reversal result is obtained.

Individually, the groups exhibit a strong, positive association.

When pooled, we observed a weak, *negative* association.

This is an example of Simpson’s Paradox . . . a reversal result.

Inclusion of relevant information is omitted in this example.

And that information would be information relative to the group.

First consider fitted values from the previous models:
Dummy Variables

Regression of Y on X2
Dummy Variables

- Simpson’s Paradox: When data from two or more groups are combined, if the direction of the relationship changes in the combined data set, then this is Simpson’s paradox. More succinctly, as Fox puts it: “that marginal and partial relationships can differ in sign is called “Simpson’s Paradox.” The marginal relationship between \( x_2 \) and \( y \) is negative but the partial relationship (controlling for \( d \)) is positive.

- Suppose we account for \( d \) in the regression?

- \( y = \beta_0 + \beta_1 X + \beta_2 d \)
# Dummy Variables

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | −0.9500  | 8.8414     | −0.11   | 0.9174   |
| x2             | 0.8250   | 0.0791     | 10.43   | 0.0000   |
| d              | 84.9500  | 5.9704     | 14.23   | 0.0000   |
Dummy Variables

- After accounting for d (or “controlling” for it as this is sometimes called), the positive relationship between $x_2$ and $y$ is retrieved.
- This model will produce parallel slopes, which is equivalent to saying, there will be two regression functions, one for $d=1$, one for $d=0$.
- Fitted function illustrates the parallel slopes result.
Dummy Variables

Regression of Y on X2

Fitted values

20 40 60 80 100 120

70 80 90 100 110 120 130 140
After accounting for \( d \) (or “controlling” for it as this is sometimes called), the positive relationship between \( x_2 \) and \( y \) is retrieved.

This model will produce parallel slopes, which is equivalent to saying, there will be two regression functions, one for \( d=1 \), one for \( d=0 \).

Fitted function illustrates the parallel slopes result.
The parallel slopes property is a function of the model being linear in the parameters.

Sometimes this may not hold.

For example, across groups, the intercept \textit{and} the slope may vary.

This may give rise to “multiplicative” models.

Important: \textit{in the general case, multiplicative models will yield much different coefficients than non-multiplicative models.}

Continue with illustration using simulated data.
## Multiplicative Terms

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 78.8394  | 1.6539     | 47.67   | 0.0000   |
| x1             | 0.9520   | 0.0518     | 18.39   | 0.0000   |
Multiplicative Terms

To this point, the linear models we have considered have all been interpreted in terms of “additive” relationships. That is, the relationship between two covariates, $X_1$ and $X_2$ in the context of the model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2,$$

is additive.

This implies that the partial effect of each independent variable is the same, regardless of the specific value at which the other covariate is held constant.

Model we estimated returns:

$$\hat{Y} = 88.81 - .766(X_1)$$
Multiplicative Terms

- $X_1$ is a quantitative covariate.
- The interpretation of the model is standard: for a 1 unit increase in $X_1$, the expected value of $Y$ decreases by about .766.
- Fitted function:
Multiplicative Terms

Model 1: $x_1$ only

![Graph showing the relationship between $x_1$ and another variable, with a trend line indicating a negative correlation.]
Now, suppose there were two distinct subpopulations related to $Y$.

To account for these two groups, we create a dummy variable called $D_1$ such that $D_1 = 1$ denotes the first group and $D_1 = 0$ denotes the second group.

Estimate model with $X_1$ and $D_1$. 
## Multiplicative Terms

|               | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 86.4269  | 4.6119     | 18.74   | 0.0000   |
| x1            | −0.6652  | 0.1458     | −4.56   | 0.0003   |
| d1            | −5.9046  | 7.9237     | −0.75   | 0.4663   |
Multiplicative Terms

- The coefficient for $D_1$ is the partial regression coefficient for the dummy variable.
- This is interpreted as giving us the expected change in $Y$ when $D_1 = 1$; that is, for group 1, the expected value of $Y$ decreases by about -5.91.
- Here, $D_1$ is going to give us an offset between the two groups.
- Must produce parallel slopes (see fitted function):
Multiplicative Terms

Model 1: x1 and d1

Group 1

Group 2
The idea of “additivity” is important here: our interpretation of $X_1$ does not change as the value of $D_1$ changes.

Similarly, our interpretation of $D_1$ does not change as a function of $X_1$.

The relationship between $D_1$ and $Y$ is accounted for by the offset in the two parallel slopes (which is equal to -5.91).

The “effects” are additive because to understand the expected value of $Y$ for both variables, we simply “add” one coefficient to the other.

The expected value of $Y$ given $D_1 = 1$ and $X_1 = 53.1$ (its mean) is $\hat{Y} = 86.43 + -0.665(53.1) - 5.91(1)$ which is 45.21.
This property ensures parallel slopes when one of the regressors is dummy variable (the slope will be offset by the value of the dummy variable).

It’s all good if the property holds.

This would imply that the effect of some covariate on the dependent variable would be influenced by changes in the value of another covariate; that is, the interpretation of, say, $X_1$ on $Y$ would be *conditional* on some other covariate, say $D_1$.

Conditional relationships, if they exist (and they often do given the kinds of theories we posit) are not immediately modeled in the additive model.
Suppose we reestimated our regression model, but this time, estimated a separate model for observations in group 1 and observations in group 2.

For group 1, our model gives us
\[ \hat{Y} = 62.11 - 0.43(X_1), \]

and for group 2, our model gives us
\[ \hat{Y} = 118.35 - 1.81(X_1). \]

For group 1, the slope coefficient tells us that for a unit change in \( X_1 \), the expected change in \( Y \) is -0.43; for group 2, that expected change is -1.81 (a little over 4 times that of group 1).

Note that if the additivity property held, the relationship between \( X_1 \) and \( Y \) would be nearly the same for the two groups. (Why?)
Multiplicative Terms

▶ To see this, return to our definition of additivity: the partial effect of each independent variable is the same, regardless of the specific value at which the other covariate is held constant.

▶ In the “two models” approach, we are essentially holding $D_1$ constant and we can see our interpretation of $X_1$ changes, depending on which value $D_1$ assumes.

▶ This is suggestive that the additivity property may not hold.

▶ Consider the fitted functions:
Multiplicative Terms

Model with $x_1$ estimated for Group 1 Only

Model with $x_1$ estimated for Group 2 Only

Model with $x_1$ (Both Groups Displayed)
Multiplicative Terms

- The point to note is how the slopes change depending on which group we are looking at.
- The slope for group 1 is considerably flatter than that for group 2 (which is consistent with our separate regression models).
- In combining the two graphs, this point is made even clearer: the slopes are not parallel and the interpretation of $X_1$ is not the same for each group.
- This suggests that the relationship between $X_1$ and $Y$ may be conditional on $D_1$; that is, $D_1$ may moderate or condition the relationship.
Multiplicative Terms

- To see this, suppose we estimated a model treating $X_1$ as a function of $D_1$. In doing this, we obtain

$$\hat{X}_1 = 28 + 50.2(D_1),$$

where the expected value of $X$ for group 1 is 78.2; the expected value of $X$ for group 2 is 28.

- The average level of $X_1$ is much higher for group 1 than group 2.

- It seems that $X_1$ is related to $D_1$.

- In terms of our additivity assumption, it will not be the case that we can correctly interpret $X_1$ by holding $D_1$ constant: the partial effect changes, depending on which group we’re looking at.
Multiplicative Terms

- This suggests that the relationship between $D_1$, $X_1$ and $Y$ is *multiplicative*.
- That is, the effect of $X_1$ on $Y$ may be multiplicatively higher in one group than when compared to the second group.
- This leads to consideration of an *interactive model*, which has the form

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 X_1 + \hat{\beta}_3 (D_1 X_1),$$

where the last term denotes the interaction term or *multiplicative term*.
- It is easy to see where the model gets its name: we are interacting, multiplicatively, $X_1$ with $D_1$. 
Multiplicative Terms

- Note what the interaction term is giving you. This term is 0 for group 2 (why?) and nonzero for group 1; hence, you’re getting an estimate of the slope of $X_1$ for $D_1$. For group 2, the relationship between $X_1$ is simply given by the coefficient for $X_1$. Thus

$$\hat{Y}_{G1} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + (\hat{\beta}_2 + \hat{\beta}_3) X_1,$$

which gives us the regression function for Group 1 and

$$\hat{Y}_{G2} = \hat{\beta}_0 + \hat{\beta}_2 X_1,$$

gives us the regression function for Group 2.

- From this model, we will not only get differences in intercepts (due to the offset between the two groups), but we will also get differences in the slope.

- For Group 1, the slope is $\hat{\beta}_2 + \hat{\beta}_3$ and for Group 2, the slope is $\hat{\beta}_2$. 
Multiplicative Terms

- Multiplicative term created by: $D_1 \times X_1$.
- Estimate model with term gives:

$$\hat{Y} = 118.35 - 56.24D_1 - 1.81X_1 + 1.38D_1X_1.$$  

- This model is not interpretable as an additive model (though it is a linear model).
- The reason is that the impact of $X_1$ on $Y$ is conditional on which group one is referencing.
- **NOTE:** sometimes analysts will refer to the coefficient for $X_1$ as a “main effect.”
- That is, sometimes the interpretation is given that the coefficient for $X_1$ represents the constant effect of $X_1$. It does not!!
## Multiplicative Terms

| Term        | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 118.3514 | 5.2817     | 22.41   | 0.0000   |
| x1          | -1.8054  | 0.1843     | -9.79   | 0.0000   |
| d1          | -56.2423 | 8.4892     | -6.63   | 0.0000   |
| x1:d1       | 1.3756   | 0.2025     | 6.79    | 0.0000   |
In the case of multiple regression, the “main effect” terminology is not justified although it is commonly assumed that the coefficients for $D_1$ and $X_1$ are “main effects.”

In the presence of an interaction, these coefficients in no instance represent a constant effect of the independent variable on the dependent variable.

The use of the term main effect implies that the coefficients are somehow interpretable alone, when they actually represent a portion of the effect of the corresponding variable on the dependent variable.

A better terminology is “constituent” terms (i.e. the constituent terms for the multiplicative term are $D_1$ and $X_1$).
Multiplicative Terms

- Model decomposition.
- For Group 1, the regression function is given by
  \[ \hat{Y} = 118.35 - 56.24D_1 + (-1.81 + 1.38)X_1, \]
  and for Group 2, the regression function is given by
  \[ \hat{Y} = 118.35 - 1.81X_1. \]
- Consider the fitted functions:
Multiplicative Terms

Interactive Model

Graph showing a decreasing trend in the Interactive Model with an x-axis variable x1 ranging from 20 to 100.
Multiplicative Terms

- Here we see nonparallel slopes.
- It should be clear that the results we obtain using the interaction term are identical to the results obtained using the separate regressions approach (refer to previous plot).
- Not coincidental!
- In the interactive model, we are conditioning the relationship of $X_1$ on $D_1$.
- This, in effect, gives us separate models within a single model.
- This is the intuition of an interactive model.
Some observations:

Recall in the model where we treated $X_1$ as a function of $D_1$ (i.e. regressed $X_1$ on $D_1$).

There, we got the expected value of $Y$ for each group.

Consider the previous plot but this time inserting a vertical reference line at the mean of $X_1$ for each group.
Multiplicative Terms
Multiplicative Terms

- The vertical reference line goes through the points 28, which corresponds to the mean of $X_1$ for Group 2 and 78.2, which corresponds to the mean of $X_1$ for Group 1.

- This figure illustrates the point made in the regression of $X_1$ on $D_1$: the separate slopes pass through the group means, which are equivalent to the regression estimates from the submodels considered.

- Extensions (but first start with the dummy variable interaction).
Posit the following model

\[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 X_1 + \hat{\beta}_3 D_1 X_1, \]

where the last term is an interaction term between the covariates \( X_1 \) and \( D_1 \).

The relationship between \( X_1 \) and \( Y \) is conditional on the presence or absence of \( D_1 \).

This gives rise to two submodels: one for group 1 and one for group 2. For group 1, the regression function is given by

\[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 1 + \hat{\beta}_2 X_1 + \hat{\beta}_3 1 X_1 = \hat{\beta}_0 + \hat{\beta}_1 + (\hat{\beta}_2 + \hat{\beta}_3) X_1. \]

and for group 2, the regression function is given by

\[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 0 + \hat{\beta}_2 X_1 + \hat{\beta}_3 0 X_1 = \hat{\beta}_0 + \hat{\beta}_2 X_1. \]
Multiplicative Terms

- For group 1, the slope is given by \((\hat{\beta}_2 + \hat{\beta}_3)\) and the intercept is given by \((\hat{\beta}_0 + \hat{\beta}_1)\).
- For group 2, the slope is given by \(\hat{\beta}_2\) and the intercept is given by \(\hat{\beta}_0\).
- In no case is the “effect” between \(X_1\) and \(Y\) constant across groups; hence there is no main effect.
- This is all easy to see in the case of a dummy variable (and even easier to see if you interact two dummy variables together); however, what about the case with two quantitative variables?
Multiplicative Terms

- Suppose we are interested in the following model,

\[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2, \]

where the two covariates are quantitative variables and the last term denotes the interaction between them.

- The intuition? The slope of \( X_1 \) on \( Y \) is conditional on \( X_2 \) and the slope of \( X_2 \) on \( Y \) is conditional on \( X_1 \).

- Since there are no main effects, the relationship of one independent variable on the dependent variable is conditional on the other independent variable.
The question arises, naturally, as to how to interpret this model?

In order to derive the conditional regression function for $Y$ and $X_1$ conditional on some specific value of $X_2$ we obtain

$$
\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2 X_2 + (\hat{\beta}_1 + \hat{\beta}_3 X_2) X_1.
$$

The conditional regression function for $Y$ and $X_2$ conditional on some specific value of $X_1$, we obtain

$$
\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + (\hat{\beta}_2 + \hat{\beta}_3 X_1) X_2.
$$
In these models, the intercept gives us the expected value of \( Y \) when all the covariates are equal to 0.

In the first model, the coefficient for \( \hat{\beta}_3 \) gives us the change in the slope of \( Y \) on \( X_1 \) associated with a unit change in \( X_2 \).

In the second model, \( \hat{\beta}_3 \) gives us the change in the slope of \( Y \) on \( X_2 \) associated with a unit change in \( X_1 \).

Key point: the slope between one covariate is governed by the other covariate.

To evaluate the impact one covariate has on the other, it is useful to note that the coefficients \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) denote the baselines around which the slopes vary.

This is easy to see (on next slide).
Multiplicative Terms

Suppose $X_2 = 0$, then the first model above reduces to

$$
\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2 0 + (\hat{\beta}_1 + \hat{\beta}_3 0) X_1 \\
= \hat{\beta}_0 + \hat{\beta}_1 X_1.
$$

Suppose $X_1 = 0$, then the second model above reduces to

$$
\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 0 + (\hat{\beta}_2 + \hat{\beta}_3 0) X_2 \\
= \hat{\beta}_0 + \hat{\beta}_2 X_2.
$$

In both cases, the models represent the “baseline” case from which the conditional slope coefficients can be evaluated.
Multiplicative Terms

- In the first model, the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ give the intercept and slope for the regression of $Y$ on $X_1$ when $X_2$ is equal to 0.
- In the second model, the coefficients $\hat{\beta}_0$ and $\hat{\beta}_2$ give the intercept and slope for the regression of $Y$ on $X_2$ when $X_1$ is equal to 0.
- Note the conditional nature of the regression model when interactions are applied.
- *In the standard least squares model without interactions, the coefficients for the covariates tell us something much different than they do in the interactive case.*
- In the interactive case, the coefficients are conditional on the value of some covariate; hence, the effect of one covariate, say $X_1$, on $Y$ is not constant across the full range of values taken by the other covariate, $X_2$. 
Multiplicative Terms

Note in passing: Suppose $X_1 = 0$, then the first model above reduces to

$$
\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2 X_2 + (\hat{\beta}_1 + \hat{\beta}_3 X_2)0 \\
= \hat{\beta}_0 + \hat{\beta}_2 X_2.
$$

Now suppose $X_2 = 0$, then the second model above reduces to

$$
\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + (\hat{\beta}_2 + \hat{\beta}_3 X_1)0 \\
= \hat{\beta}_0 + \hat{\beta}_1 X_1.
$$

Implications??

The reported regression coefficient $\hat{\beta}_1$ and $\hat{\beta}_2$ only has a unique interpretation when $X_2$ or $X_1$ is 0.
Because of this conditional property, coefficient estimates from an interactive model will in general be much different than estimates derived from a model possessing the additivity property.

Note also, the importance of “0” in model decomposition.

The standard error for the reported constituent term will only make sense at the point at which the conditioning variable assumes the value “0” (why?).

If the $X_k$ (constituent variables) do not have a natural 0 point, then the reported standard errors on the constituent terms are utterly irrelevant.

Implications?
Multiplicative Terms

- Having a 0 point was convenient because it gave the coefficients a natural interpretation: the baseline against which changes in the conditional slope could be evaluated.
- In the absence of a meaningful 0 term, the baseline “effects” are never observed in the data.
- Hence, the coefficient estimate (which is interpreted in terms of the covariate having a 0 point) and its standard error can be misleading, sometimes very misleading, because it is giving you the conditional relationship under a condition that doesn’t actually hold in the data.
- **Common Mistake**: Analysts conclude constituent term parameters are “insignificant” based solely on inspecting regression output.
- A small t-value at the point of 0 on the conditioning variable does not mean the t will be larger over other ranges of the conditioning variable.
Some observations: proper interpretation of the multiplicative regression model requires a lot more work.

To see why, motivate it with the following model:

\[ Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon \]

(some of what follows comes from Brambor et al [2006]).

Z may be discrete or continuous.

Suppose Z is continuous for the moment.

Useful to consider “two-sided conditioning” in the multiplicative model.
In our model, $\beta_1$ gives the one-unit change in $X$ on the expected value of $Y$ only when $Z = 0$.

That is: $\left( \frac{\partial(Y)}{\partial(X_1)} \mid Z = 0 \right) = \beta_1$

Simplifying:

$$Y = \beta_0 + \beta_2 Z + (\beta_1 + \beta_3 Z)X_1$$

If $Z$ is a binary variable, the simplification is more straightforward:

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1$$

noting that $Z$ is implied to be 1.

In either model, $\beta_1$ only gives the “average effect” of $X_1$ on $Y$ when $Z = 0$.

Also in either model, the relationship between $X$ and $Y$ is conditioned by $Z$. 

The “flip-side”: $\beta_2$ gives the one-unit change in $Z$ on the expected value of $Y$ only when $X_1 = 0$.

That is: \[ \left( \frac{\partial(Y)}{\partial(Z)} \bigg| X_1 = 0 \right) = \beta_2 \]

Simplifying:

\[ Y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1)Z \]

Again: $\beta_1$ only gives the “average effect” of $X_1$ on $Y$ when $Z = 0$.

Also: the relationship between $Z$ and $Y$ is conditioned by $X_1$. 
Multiplicative Terms

- A multiplicative hypothesis: (language from Brambor et al) "An increase in $X$ is associated with an increase in $Y$ when $Z$ is present, but not when $Z$ is absent."

- More general: "An increase in $X$ is associated with an increase in $Y$ as $Z$ departs from 0."

- In other words, the relationship between $X$ and $Y$ is conditional on the value $Z$ assumes.

- If both the constituent terms are continuous, the model is "fully" conditional meaning we cannot say anything about $X_1$ (or $Z$) without knowing something about $Z$ (or $X_1$).

- **Common Mistake:** Analysts often proceed as giving an unconditional interpretation. This cannot be sustained by the model that is posited.
Multiplicative Terms

- Estimating a model (Quick illustration)
- Let $X_1$ and $X_2$ be two quantitative variables.
- Estimate: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$
- Using the simulated data, we return:

$$\hat{Y} = 6.43 + .18X_1 + .99X_2 - .011X_1X_2.$$
## Multiplicative Terms

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 6.4296   | 5.4078     | 1.19    | 0.2518   |
| x1         | 0.1770   | 0.0586     | 3.02    | 0.0081   |
| x2         | 0.9860   | 0.0527     | 18.70   | 0.0000   |
| x1:x2      | -0.0108  | 0.0014     | -7.53   | 0.0000   |
Multiplicative Terms

- Question: what do the parameters for $X_1$ and $X_2$ mean?
- Question: suppose either of these parameters are no different from 0?
- Question: suppose one is and one is not?
- We will return to these questions in a bit.
- Back to some quick interpretation.
Multiplicative Terms

Because the relationship between $X_1$ and $Y$ will be conditional on $X_2$ (and vise versa), we will want to consider “conditional slope” coefficients.

That is, the slope varies over the range of the conditioning variable.

Easy to see if we rearrange our model: the conditional slope (the term in parentheses) for $Y$ regressed on $X_1$ conditional on $X_2$ is

$$\hat{Y} = 6.43 + .99X_2 + (.18 - .011X_2)X_1.$$ 

and the conditional slope for $Y$ regressed on $X_2$ conditional on $X_1$ is

$$\hat{Y} = 6.43 + .18X_1 + (.99 - .011X_1)X_2.$$
Multiplicative Terms

- It should be clear that the “average effect” for $X_1$ on $Y$ is never unconditionally related to $X_2$. *This is a fundamental difference between this and the additive model.*
- For $X_1$, the range is 21 to 100; for $X_2$, the range is 40 to 100.
- Suppose we compute the predicted regression function of the slope of $Y$ on $X_1$ for the maximum and minimum and mean values of $X_2$?
- Plot the predicted functions (next slide).
Multiplicative Terms

Interactive Model: \( x_2 = 40, 65, 100 \)

Slope of \( Y \) on \( X_1 \) Conditional on \( X_2 \)
Multiplicative Terms

- Note is the slope of the three lines, which represent hypothetical regression functions.
- These lines are created by substituting 40, 65, and 100 into the first model above and then allowing $X_1$ to freely vary.
- This gives us a sense of the conditional relationship and it tells us that as $X_2$ decreases, the slope between $Y$ and $X_1$ tends to decrease, that is, flatten out.
- To see this numerically, note that the estimated slope for $X_2 = 100$ is about -.92; for $X_2 = 40$, the estimated slope is about -.26. The dots in the graph represent the actual predicted values of $Y$ given the actual values of $X_1$ and $X_2$. 
Multiplicative Terms

- Repeat the same exercise for the slope of $Y$ on $X_2$.
- The conclusion here is that as $X_1$ increases, the slope between $Y$ and $X_2$ tends to decrease (it becomes negative for the upper bound on $X_1$).
- The range of the estimated slopes is $-0.11$ (for $X_1 = 100$) to $0.76$ (for $X_1 = 21$).
- Plot the function. Again, the dots represent the actual predicted values of $Y$ given the actual values of $X_1$ and $X_2$.
- Important! I’ve used the phrase “two-sided” conditioning but there is obviously only 1 predicted value for any coordinate $(X_1, X_2)$. 
Multiplicative Terms

Interactive Model: x1=21, 53, 100

Slope of Y on X2 Conditional on X1
Multiplicative Terms

- The point of this exercise is to note that inclusion of, and interpretation of interaction terms is straightforward.
- Yet people mess this up all the time.
- There are caveats: you need to be careful in interpreting the results.
- Plots like the ones shown above are problematic because many (most!) of the points on the predicted lines are not actual data points—they are hypothetical.
- Side-trip into problem-land.
Multiplicative Terms

Return to a model with a dummy variable $D$ and a quantitative variable $X$.

What follows is generalizable to all settings.

Suppose we estimate a model with $X_1D_1$ included as a predictor?

Further, imagine the regression coefficient for $D_1$ is no different from 0.

Thus instead of $Y = \beta_0 + \beta_1X_1 + \beta_2D_1 + \beta_3X_1D_1$ we estimate $Y = \beta_0 + \beta_1X_1 + \beta_3X_1D_1$. 
Multiplicative Terms

- Rearranging, we fit: \( Y = \beta_0 + \beta_1 X_1 \iff D_1 = 0 \) and 
  \( Y = \beta_0 + (\beta_1 X_1 + \beta_3)X_1 \iff D_1 = 1 \).
- This approach is not infrequently taken in the literature.
- BUT, it is almost always a bad idea!
- Consider the fitted function (using the simulated data and variables \( X_1 \) and \( D_1 \)).
## Multiplicative Terms

| Term     | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 96.5801  | 7.7613     | 12.44   | 0.0000   |
| x1       | -1.0629  | 0.2747     | -3.87   | 0.0012   |
| x1:d1    | 0.2049   | 0.1855     | 1.10    | 0.2848   |
Multiplicative Terms

Omitting the Parm for D1

Graph showing the relationship between two variables, with lines labeled pval2 and pval1.
Multiplicative Terms

- Implications?
- Omitting $\beta_2$ forces the regression lines to meet at a common $y$-intercept.
- This is an awkward and probably unnatural functional form: two groups start at the same place and then diverge.
- Also makes the strong assumption that you unequivocally know $\beta_2 = 0$ in the population.
- Suppose now you keep $\beta_2$ but omit $\beta_1$?
## Multiplicative Terms

|         | Estimate | Std. Error | t value | Pr(>|t|) |
|---------|----------|------------|---------|----------|
| (Intercept) | 67.8000  | 2.8771     | 23.57   | 0.0000   |
| d1      | -5.6909  | 17.2942    | -0.33   | 0.7461   |
| d1:x1   | -0.4298  | 0.2149     | -2.00   | 0.0618   |
Multiplicative Terms

Omitting Parm for X1
We now have an equally awkward functional form: when $D_1 = 1$, the slope is non-zero; when $D_1 = 0$, the slope is 0.

Again note: this is NOT a function of the “real world” but a function of the model’s specification.

You are assuming you know that in the population there is no relationship between $X_1$ and $Y$ when condition $D_1$ is absent.

This is a strong assumption and one not usually justified.

Despite temptation to do otherwise, always include the constituent terms.
Multiplicative Terms

- The marginal “effect” or conditional slope is \((\beta_1 + \beta_3 X_2)\).
- It is clear then that the relationship between \(Y\) and \(X_1\) is going to be conditioned by the level of \(X_2\).
- If the slope varies, the standard error must vary as well.
- This means the following: simply reporting coefficients from a multiplicative model (especially one where the two constituent terms are quantitative variables) will be inadequate.
- Much more work is needed.
Multiplicative Terms

- Understand that the standard error reported to you is the standard error for the case when $X_1$ and $X_2$, respectively, are zero.

- That is, *the standard errors are only applicable to a particular range of the interactive effect.*

- It is not at all uncommon to estimate a statistical interaction and find that the standard errors, relative to the coefficient, are very large, and indeed, statistically insignificant. This insignificance can be misleading (although it won’t always be misleading).

- Just as the relationship between $X_1$ and $Y$ in an interactive model is conditional on $X_2$, *so to are the standard errors.*
Multiplicative Terms

Because multiple terms are involved in the multiplicative effect, the standard error must be modified to account for the variances and covariances of the $\beta_1$ and $\beta_3$.

The general formula to compute the standard errors of the conditional slopes is given by

$$s.e.\hat{\beta}_{1+3} = \sqrt{\text{var}(\hat{\beta}_1) + X_2^2 \text{var}(\hat{\beta}_3) + 2X_2 \text{cov}(\hat{\beta}_1 \hat{\beta}_3)},$$

for the slope of $Y$ on $X_1$ conditional on $X_2$.

This may also be notated as $\sigma_{\frac{\partial Y}{\partial X}}$.

Note also that:

$$s.e.\hat{\beta}_{2+3} = \sqrt{\text{var}(\hat{\beta}_2) + X_1^2 \text{var}(\hat{\beta}_3) + 2X_1 \text{cov}(\hat{\beta}_2 \hat{\beta}_3)},$$

is the slope of $Y$ on $X_2$ conditional on $X_1$. 
Multiplicative Terms

- Note that this formula is not a constant; it is determined by the level of the covariate upon which the slope is conditioned. In the first case, this is $X_2$; in the second case, this is $X_1$.
- Illustration using the simulated data.
- Recall that we estimated the following model:

$$\hat{Y} = 6.43 + .18X_1 + .99X_2 - .011X_1X_2.$$ 

- The conditional slope for $Y$ regressed on $X_1$ conditional on $X_2$ was given by

$$\hat{Y} = 6.43 + .99X_2 + (.18 - .011X_2)X_1,$$

and the conditional slope for $Y$ regressed on $X_2$ conditional on $X_1$ was given by

$$\hat{Y} = 6.43 + .18X_1 + (.99 - .011X_1)X_2.$$
The quantity of primary interest is the term in the last two equations in parentheses.

This gives us the conditional slope or marginal effect.

What we want to do is compute the standard error for these conditional slopes, which will vary as $X_2$ (or $X_1$) varies.

Looking at the first submodel, suppose that $X_2$ is set at 100.

The conditional slope is about -.90 (rounding).

The question to ask is whether or not this conditional slope coefficient is different from 0?
Multiplicative Terms

▸ To answer this question, we need to compute the standard error.

▸ In order to compute the standard error, we need to back out of our coefficient estimates, the variances and covariances of the parameters.

▸ In R use the syntax: `vcov(lm object)`: 

```r
> vcov(idmod4)

     (Intercept)       x1       x2       x1:x2
(Intercept) 29.24397  -0.11652  -0.26088  -0.00192
  x1        0.11652  3.43021   2.07824  -0.00192
  x2        0.26088  2.07824   2.77936  -0.00192
x1:x2      -0.00192  0.00691  -0.10659  0.00042
```
Using elements from this matrix, obtain the following variances and covariances that can be substituted into the formula above.

For the slope of $Y$ on $X_1$ conditional on $X_2 = 100$:

$$s.e.\hat{\beta}_1 + \hat{\beta}_3 = \sqrt{.0034 + 100^2(.00002042) + 2(100)(-.000067)},$$

This returns a standard error estimate of about $.102$. Given the conditional slope estimate for this case was about -.92, I see that the $t$ ratio is about -9, which would be significant at any level.

Visualize this: I plot the conditional slope estimates as well as the upper and lower 95 percent confidence limits around the estimate.
Multiplicative Terms

Slope of Y on X1, Conditional on X2

Conditional Slope Estimates with 95 Percent Confidence Interval
Multiplicative Terms

Slope of Y on X2, Conditional on X1

Conditional Slope Estimates with 95 Percent Confidence Interval
Multiplicative Terms

- As can be seen, the slope of $Y$ on $X_1$ conditional on $X_2$ is in general more precisely estimated than the slope of $Y$ on $X_2$, conditional on $X_1$.
- The major point to understand here is that if you are going to interpret these interactions, you need to be aware that the range of effects are not all going to be statistically different from 0.
- To illustrate these points further, we could compute the $t$ ratios for each of the predicted slope coefficients.
- This could be done easily by deriving the conditional slope and dividing it by its standard error.
- As these are $t$ ratios, any $t$ exceeding the critical $t$ would be statistically significant.
- For these data, I set $\alpha = .05$ and compute the two-tail critical $t$, which is 2.12.
Multiplicative Terms

Estimated t-ratios for Conditional Slopes

- Estimated t-ratios for Conditional Slopes

40 60 80 100

x2
Multiplicative Terms

Estimated t-ratios for Conditional Slopes

- x1

- 0

- 20

- 40

- 60

- 80

- 100

- 0

- 5

- 10

- 15
Multiplicative Terms

- In the first figure, I find that all of the estimated $t$ ratios exceed the critical $t$ of 2.12.
- This implies that the conditional slopes are statistically different from 0, and implies that across the full range of $X_2$, the interaction effect seems to be significant.
- In the second figure, we see that once $X_1$ exceeds 75 (or so), the estimated $t$ ratios drop below the critical threshold (which is denoted by the horizontal reference line at 2.12).
- *For this range of data, the statistical interaction is not different from 0.*
- This is illustrative of the larger point that the statistical interaction need not be significant across the full range of the conditioning variable, in this case, $X_1$. 
Multiplicative Terms

- Useful resource: Matt Golder (soon-to-be at PSU) but currently at Florida State has very useful Stata modules to compute marginal effects and standard errors.