Chapter Four: Multiple Regression II

Interactions as Products of Independent Variables

Categorical by Quantitative

An interaction between a quantitative variable and a categorical variable means that differences in E[Y] between categories depend on the value of the quantitative variable, or (equivalently) that the slope of the lines relating x to E[Y] are different, depending on category membership. Such an interaction is represented by **products** of the quantitative variable and the dummy variables for the categorical variable.

For example, consider the metric cars data (mcars.dat). It has length, weight, origin and fuel efficiency in kilometers per litre, for a sample of cars. The three origins are US, Japanese and Other. Presumably these refer to the location of the head office, not to where the car was manufactured.

Let's use indicator dummy variable coding for origin, with an intercept. In an Analysis of Covariance (ANCOVA), we'd test country of origin controlling, say, for weight. Letting x represent weight and c1 and c2 the dummy variables for country of origin, the model would be

$$E[Y|\mathbf{X}] = b_0 + b_1x + b_2c_1 + b_3c_2.$$

This model assumes no interaction between country and weight. The following model includes product terms for the interaction, and would allow you to test it.

$$E[Y|\mathbf{X}] = \beta_0 + \beta_1x + \beta_2c_1 + \beta_3c_2 + \beta_4c_1x + \beta_5c_2x$$

Country	c1	c2	Expected KPL (let x = weight)
U. S.	1	0	$(\beta_0 + \beta_2) + (\beta_1+\beta_4)x$
Japan	0	0	$\beta_0 + \beta_1 x$
European	0	1	$(\beta_0 + \beta_3) + (\beta_1+\beta_5)x$

It's clear that the slopes are parallel if and only if $\beta_4=\beta_5=0$, and that in this case the relationship of fuel efficiency to country would not depend on weight of the car.

As the program below shows, interaction terms are created by literally multiplying independent variables, and using products as additional independent variables in the regression equation.

```
/********************** mcars.sas **************************/
options linesize=79 pagesize=100 noovp formdlim='-';
title 'Metric Cars Data: Dummy Vars and Interactions';

proc format; /* Used to label values of the categorical variables */
     value carfmt    1 = 'US'
                     2 = 'Japanese'
                     3 = 'European' ;
data auto;
     infile 'mcars.dat';
     input id country kpl weight length;
/* Indicator dummy vars: Ref category is Japanese */
     if country = 1 then c1=1;  else c1=0;
     if country = 3 then c2=1;  else c2=0;
     /* Interaction Terms */
     cw1 = c1*weight; cw2 = c2*weight;
     label country = 'Country of Origin'
           kpl = 'Kilometers per Litre';
     format country carfmt.;

proc means;
     class country;
     var weight kpl;

proc glm;
     title 'One-way ANOVA';
     class country;
     model kpl = country;
     means country / tukey;

proc reg;
     title 'ANCOVA';
     model kpl = weight c1 c2;
     country: test c1 = c2 = 0;
```

```
proc reg;
     title 'Test parallel slopes (Interaction)';
     model kpl = weight c1 c2 cw1 cw2;
     interac: test cw1 = cw2 = 0;
     useuro:  test cw1=cw2;
     country: test c1 = c2 = 0;
     eqreg:   test c1=c2=cw1=cw2=0;

proc iml; /* Critical value for Scheffe tests */
     critval = finv(.95,4,94) ; print critval;



/* Could do most of it with proc glm: ANCOVA, then test interaction */

proc glm;
     class country;
     model kpl = weight country;
     lsmeans country;

proc glm;
     class country;
     model kpl = weight country weight*country;
```

Let's take a look at the output. First, proc means indicates that the US cars get lower gas mileage, and that weight is a potential confounding variable.

```
       COUNTRY   N Obs  Variable  Label                    N          Mean
      ------------------------------------------------------------------
      US           73   WEIGHT                            73       1540.23
                        KPL       Kilometers per Litre    73     8.1583562

      Japanese     13   WEIGHT                            13       1060.27
                        KPL       Kilometers per Litre    13     9.8215385

      European     14   WEIGHT                            14       1080.32
                        KPL       Kilometers per Litre    14    11.1600000
      ------------------------------------------------------------------

   COUNTRY  N Obs  Variable  Label                         Std Dev       Minimum
  ---------------------------------------------------------------------------
  US           73  WEIGHT                              327.7785402   949.5000000
                   KPL       Kilometers per Litre        1.9760813     5.0400000

  Japanese     13  WEIGHT                              104.8370989   891.0000000
                   KPL       Kilometers per Litre        2.3976719     7.5600000

  European     14  WEIGHT                              240.9106607   823.5000000
                   KPL       Kilometers per Litre        4.2440764     5.8800000
```

```
 ---------------------------------------------------------------------------

        COUNTRY  N Obs  Variable  Label                        Maximum
       -------------------------------------------------------------
       US           73  WEIGHT                                 2178.00
                        KPL       Kilometers per Litre      12.6000000

       Japanese     13  WEIGHT                                 1237.50
                        KPL       Kilometers per Litre      14.7000000

       European     14  WEIGHT                                 1539.00
                        KPL       Kilometers per Litre      17.2200000
       -------------------------------------------------------------
```

The one-way ANOVA indicates that fuel efficiency is significantly related to country of origin; country explains 17% of the variation in fuel efficiency.

```
                           General Linear Models Procedure

   Dependent Variable: KPL   Kilometers per Litre
                                           Sum of            Mean
   Source                  DF             Squares          Square  F Value    Pr > F

   Model                    2     121.59232403     60.79616201    10.09    0.0001
   Error                   97     584.29697197      6.02368012
   Corrected Total         99     705.88929600

                     R-Square             C.V.        Root MSE           KPL Mean
                     0.172254         27.90648       2.4543187          8.7948000
```

The Tukey follow-ups are not shown, but they indicate that only the US-European difference is significant. Maybe the US cars are less efficient because they are big and heavy. So let's do the same test, controlling for weight of car. Here's the SAS code. Note this is a standard Analysis of Covariance, and we're *assuming* no interaction.

```
proc reg;
     title 'ANCOVA';
     model kpl = weight c1 c2;
     country: test c1 = c2 = 0;
```

```
   Dependent Variable: KPL        Kilometers per Litre

                                Analysis of Variance
                                Sum of         Mean
       Source          DF      Squares       Square      F Value       Prob>F

       Model            3    436.21151    145.40384       51.761       0.0001
       Error           96    269.67779      2.80914
       C Total         99    705.88930

           Root MSE       1.67605     R-square       0.6180
           Dep Mean       8.79480     Adj R-sq       0.6060
           C.V.          19.05728

                              Parameter Estimates

                      Parameter      Standard    T for H0:
     Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

     INTERCEP   1     16.226336    0.76312281        21.263        0.0001
     WEIGHT     1     -0.006041    0.00057080       -10.583        0.0001
     C1         1      1.236147    0.57412989         2.153        0.0338
     C2         1      1.459591    0.64565633         2.261        0.0260

-------------------------------------------------------------------------------

Dependent Variable: KPL
Test: COUNTRY  Numerator:      8.6168  DF:    2  F value:   3.0674
               Denominator:  2.809144  DF:   96  Prob>F:    0.0511
```

First notice that by including weight, we're now explaining 61% of the variation, while before we explained just 17%. Also, while the effect for country was comfortably significant before we controlled for weight, now it narrowly fails to reach the traditional criterion (p = `0.0511`). But to really appreciate these results, we need to make a table.

Country	c1	c2	$E[Y] = \beta_0 + \beta_1 x + \beta_2 c_1 + \beta_3 c_2$
U. S.	1	0	$(\beta_0 + \beta_2) + \beta_1 x$
Japan	0	0	$\beta_0 \qquad + \beta_1 x$
European	0	1	$(\beta_0 + \beta_3) + \beta_1 x$

```
                              Parameter Estimates

                      Parameter      Standard    T for H0:
     Variable  DF      Estimate         Error   Parameter=0    Prob > |T|
     INTERCEP   1      16.226336    0.76312281        21.263        0.0001
     WEIGHT     1      -0.006041    0.00057080       -10.583        0.0001
     C1         1       1.236147    0.57412989         2.153        0.0338
     C2         1       1.459591    0.64565633         2.261        0.0260
```

Observe that both b_2 and b_3 are positive -- and significant. Before we controlled for weight, Japanese gas mileage was a little better than US, though not significantly so. Now, because b_2 estimates β_2, and β_2 is the population difference between U.S. and Japanese mileage (for any fixed weight), a positive value of b_2 means that once you control for weight, the U.S. cars are getting better gas mileage than the Japanese -- significantly better, too, if you believe the t-test and not the F-test.

The *direction* of the results has changed because we controlled for weight. This can happen.

Also, may seem strange that the tests for β_2 and β_3 are each significant individually, but the simultaneous test for both of them is not. But this the simultaneous test implicitly includes a comparison between U.S. and European cars, and they are very close, once you control for weight.

The best way to summarize these results would be to calculate Y-hat for each country of origin, with weight set equal to its mean value in the sample. Instead of doing that, though, let's first test the interaction, which this analysis is *assuming* to be absent.

```
proc reg;
     title 'Test parallel slopes (Interaction)';
     model kpl = weight c1 c2 cw1 cw2;
        interac: test cw1 = cw2 = 0;
        useuro:  test cw1=cw2;
        country: test c1 = c2 = 0;
        eqreg:   test c1=c2=cw1=cw2=0;
```

```
   Dependent Variable: KPL        Kilometers per Litre

                                    Sum of         Mean
          Source           DF      Squares       Square       F Value       Prob>F

          Model             5    489.27223     97.85445        42.463       0.0001
          Error            94    216.61706      2.30444
          C Total          99    705.88930

              Root MSE       1.51804     R-square       0.6931
              Dep Mean       8.79480     Adj R-sq       0.6768
              C.V.          17.26062
```

```
                              Parameter Estimates

                        Parameter      Standard    T for H0:
     Variable  DF        Estimate         Error   Parameter=0    Prob > |T|

     INTERCEP   1       29.194817    4.45188417         6.558        0.0001
     WEIGHT     1       -0.018272    0.00418000        -4.371        0.0001
     C1         1      -12.973668    4.53404398        -2.861        0.0052
     C2         1       -4.891978    4.85268101        -1.008        0.3160
     CW1        1        0.013037    0.00421549         3.093        0.0026
     CW2        1        0.006106    0.00453064         1.348        0.1810

 ---------------------------------------------------------------------------------
```

```
Dependent Variable: KPL
Test: INTERAC  Numerator:      26.5304  DF:    2   F value:  11.5127
               Denominator:  2.304437  DF:   94   Prob>F:    0.0001

Dependent Variable: KPL
Test: USEURO   Numerator:      33.0228  DF:    1   F value:  14.3301
               Denominator:  2.304437  DF:   94   Prob>F:    0.0003

Dependent Variable: KPL
Test: COUNTRY  Numerator:      24.4819  DF:    2   F value:  10.6238
               Denominator:  2.304437  DF:   94   Prob>F:    0.0001

Dependent Variable: KPL
Test: EQREG    Numerator:      17.5736  DF:    4   F value:   7.6260
               Denominator:  2.304437  DF:   94   Prob>F:    0.0001
```

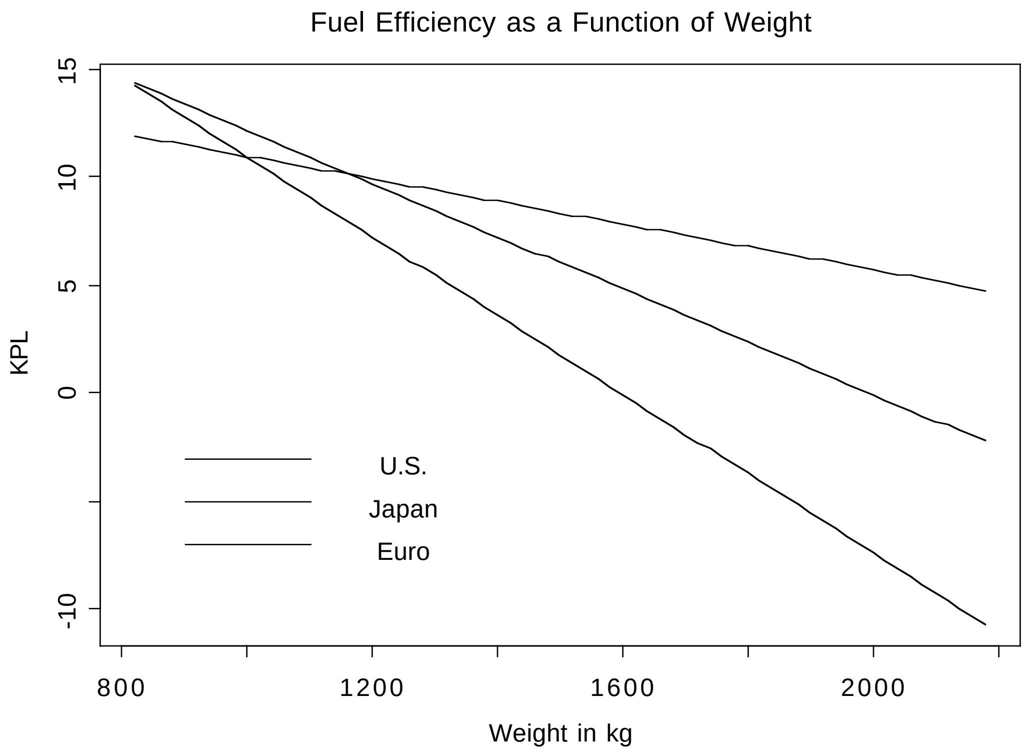
Now the coefficients for the dummy variables are both negative, and the coefficients for the interaction terms are positive. To see what's going on, we need a table *and* a picture -- of $\hat{Y}$.

$$\hat{Y} = b_0 + b_1x + b_2c_1 + b_3c_2 + b_4c_1x + b_5c_2x$$

$$= 29.194817 - 0.018272x - 12.973668c_1 - 4.891978c_2 + 0.013037c_1x + 0.006106c_2x$$

Country	c1	c2	Predicted KPL (let x = weight)
U. S.	1	0	$(b_0 + b_2) + (b_1+b_4)x$ = 16.22 - 0.005235 x
Japan	0	0	$b_0 + b_1 x$ = 29.19 - 0.018272 x
European	0	1	$(b_0 + b_3) + (b_1+b_5)x$ = 24.30 - 0.012166 x

From the proc means output, we find that the lightest car was 823.5kg, while the heaviest was 2178kg. So we will let the graph range from 820 to 2180.



When there were no interaction terms, b2 and b3 represented a main effect for country. What do they represent now?

From the picture, it is clear that the most interesting thing is that the slope of the line relating weight to fuel efficiency is least steep for the U.S. Is it significant? 0.05/3 = 0.0167.

Repeating earlier material, ...

```
                              Parameter Estimates

                       Parameter      Standard    T for H0:
        Variable  DF    Estimate         Error   Parameter=0    Prob > |T|

        INTERCEP   1    29.194817    4.45188417        6.558        0.0001
        WEIGHT     1    -0.018272    0.00418000       -4.371        0.0001
        C1         1   -12.973668    4.53404398       -2.861        0.0052
        C2         1    -4.891978    4.85268101       -1.008        0.3160
        CW1        1     0.013037    0.00421549        3.093        0.0026
        CW2        1     0.006106    0.00453064        1.348        0.1810



        useuro:  test cw1=cw2;
```

```
Dependent Variable: KPL
Test: USEURO   Numerator:     33.0228  DF:    1   F value:  14.3301
               Denominator:  2.304437  DF:   94   Prob>F:    0.0003
```

The conclusion is that with a Bonferroni correction, the slope is less (less steep) for US than for either Japanese or European, but Japanese and European are not significantly different from each other.

Another interesting follow-up would be to use Scheffé tests to compare the heights of the regression lines at many values of weight; infinitely many comparisons would be protected simultaneously. This is not a proper follow-up to the interaction. What is the initial test?

Quantitative by Quantitative

An interaction of two quantitative variables is literally represented by their product. For example, consider the model

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Hold x_2 fixed at some particular value, and re-arrange the terms. This yields

$$E[Y] = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2)x_1.$$

so that there is a linear relationship between x_1 and E[Y], with both the slope and the intercept depending on the value of x_2. Similarly, for a fixed value of x_1,

$$E[Y] = (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1)x_2,$$

and the (linear) relationship of x_2 to E[Y] depends on the value of x_1. We always have this kind of symmetry.

Three-way interactions are represented by 3-way products, etc. Its interpretation would be "the 2-way interaction depends ..."

Product terms represent interactions ONLY when all the variables involved and all lower order interactions involving those variables are also included in the model!

Categorical by Categorical

It is no surprise that interactions between categorical independent variables are represented by products. If A and B are categorical variables, IVs representing the A by B interaction are obtained by multiplying each dummy variable for A by each dummy variable for B. If there is a third IV cleverly named C and you want the 3-way interaction, multiply each of the dummy variables for C by each of the products representing the A by B interaction. This rule extends to interactions of any order.

Up till now, we have represented categorical independent variables with indicator dummy variables, coded 0 or 1. If interactions between categorical IVs are to be represented, it is much better to use "effect coding," so that the regression coefficients for the dummy variables correspond to main effects. (In a 2-way design, products of indicator dummy variables still correspond to interaction terms, but if an interaction is present, the interpretation of the coefficients for the indicator dummy variables is not what you might guess.)

Effect coding. There is an intercept. As usual, a categorical independent variable with k categories is represented by k-1 dummy variables. The rule is

Dummy var 1: First value of the IV gets a 1, last gets a minus 1, all others get zero.
Dummy var 2: Second value of the IV gets a 1, last gets a minus 1, all others get zero.
. . .
Dummy var k-1: k-1st value of the IV gets a 1, last gets a minus 1, all others get zero.

In the Greenhouse data, there are six genetically different types of fungus growing on three varieties of Canola plant. The dependent variablle is lesion length -- how hig a wound the fungus made on the plant after ten days. Here is a table showing effect coding for Plant.

Plant	p1	p2	$E[Y\|\mathbf{X}] = \beta_0 + \beta_1 p_1 + \beta_2 p_2$
GP159	1	0	$\mu_1 = \beta_0 + \beta_1$
Hanna	0	1	$\mu_2 = \beta_0 + \beta_2$
Westar	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

It is clear that $\mu_1 = \mu_2 = \mu_3$ if and only if $\beta_1=\beta_2=0$, so it's a valid dummy variable coding scheme even though it looks strange.

Country	p1	p2	$E[Y] = \beta_0 + \beta_1 p_1 + \beta_2 p_2$
GP159	1	0	$\mu_1 = \beta_0 + \beta_1$
Hanna	0	1	$\mu_2 = \beta_0 + \beta_2$
Westar	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

Effect coding has these properties, which extend to any number of categories.

- $\mu_1 = \mu_2 = \mu_3$ if and only if $\beta_1=\beta_2=0$.
- The average population mean (grand mean) is $(\mu_1+\mu_2+\mu_3)/3 = \beta_0$.
- β_1, β_2 and $-(\beta_1+\beta_2)$ are deviations from the grand mean.

The real advantage of effect coding is that the dummy variables behave nicely when multiplied together, so that main effects correspond to collections of dummy variables, and interactions correspond to their products -- in a simple way. This is illustrated for Plant by Fungus Type. Fungus type is called MCG for "Mycelial Compatibility Group." This strange name comes from the way that the botanists decided whether two types of fungus were genetically distinct. They would grow two samples on the same dish in a nutrient solution, and if the two fungus patches stayed separate, they were genetically different. If they grew together into a single patch of fungus (that is, they were compatible), then they were genetically identical. Apparently, this phenomenon is well established.

```
data nasty;
     set yucky;
     /* Two dummy variables for plant */
        if plant=. then p1=.;
        else if plant=1 then p1=1;
        else if plant=3 then p1=-1;
        else p1=0;
```

```
if plant=. then p2=.;
   else if plant=2 then p2=1;
   else if plant=3 then p2=-1;
   else p2=0;
/* Five dummy variables for mcg */
if mcg=. then f1=.;
   else if mcg=1 then f1=1;
   else if mcg=9 then f1=-1;
   else f1=0;
if mcg=. then f2=.;
   else if mcg=2 then f2=1;
   else if mcg=9 then f2=-1;
   else f2=0;
if mcg=. then f3=.;
   else if mcg=3 then f3=1;
   else if mcg=9 then f3=-1;
   else f3=0;
if mcg=. then f4=.;
   else if mcg=7 then f4=1;
   else if mcg=9 then f4=-1;
   else f4=0;
if mcg=. then f5=.;
   else if mcg=8 then f5=1;
   else if mcg=9 then f5=-1;
   else f5=0;
/* Product terms for the interaction */
   p1f1 = p1*f1; p1f2=p1*f2 ; p1f3=p1*f3 ; p1f4=p1*f4; p1f5=p1*f5;
   p2f1 = p2*f1; p2f2=p2*f2 ; p2f3=p2*f3 ; p2f4=p2*f4; p2f5=p2*f5;
```

```
proc reg;
     model meanlng = p1 -- p2f5;
     plant:  test p1=p2=0;
     mcg:    test f1=f2=f3=f4=f5=0;
     p_by_f: test p1f1=p1f2=p1f3=p1f4=p1f5=p2f1=p2f2=p2f3=p2f4=p2f5 = 0;
```

Here is the output from the test statement. For comparison, it is followed by `proc glm` output from `model meanlng = plant|mcg` (a standard two-way ANOVA).

```
Dependent Variable: MEANLNG
Test: PLANT    Numerator: 110847.5637  DF:    2   F value: 113.9032
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: MCG      Numerator:  11748.0529  DF:    5   F value:  12.0719
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: P_BY_F   Numerator:   4758.1481  DF:   10   F value:   4.8893
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001

-------------------------------------------------------------------------------


Source                  DF       Type III SS     Mean Square  F Value     Pr > F

PLANT                    2      221695.12747    110847.56373   113.90     0.0001
MCG                      5       58740.26456     11748.05291    12.07     0.0001
PLANT*MCG               10       47581.48147      4758.14815     4.89     0.0001
```

It worked.

Effect coding works as expected in conjunction with quantitative independent variables. In particular, products of quantitative and indicator variables still represent interactions. In fact, the big advantage of effect coding is that you can use it to test categorical independent variables, and interactions between categorical independent variables -- in a bigger multiple regression context.