# LECTURE 4 - CORRELATION & REGRESSION – PART D

Regression with Interaction among Predictors

# Multiple Regression – Interaction

- Sometimes, *predictors* in a multiple regression model *interact* with each other *in determining the predicted value* of the dependent variable Y. We call such models "interaction models".

- Visually, the interaction between two predictors can be captured by plotting the value of Y (or the predicted value of Y) against one of the predictors, for different values of the other predictors.

- If the slope (**β**) of a predictor is significantly different for different values of the other predictor, then we say we have significant interaction.

# INTERACTION BETWEEN CONTINUOUS AND CATEGORICAL PREDICTORS

Lecture 4D-1

# Multiple Regression – Interaction between Categorical & Continuous Predictors (**Interact.R**)

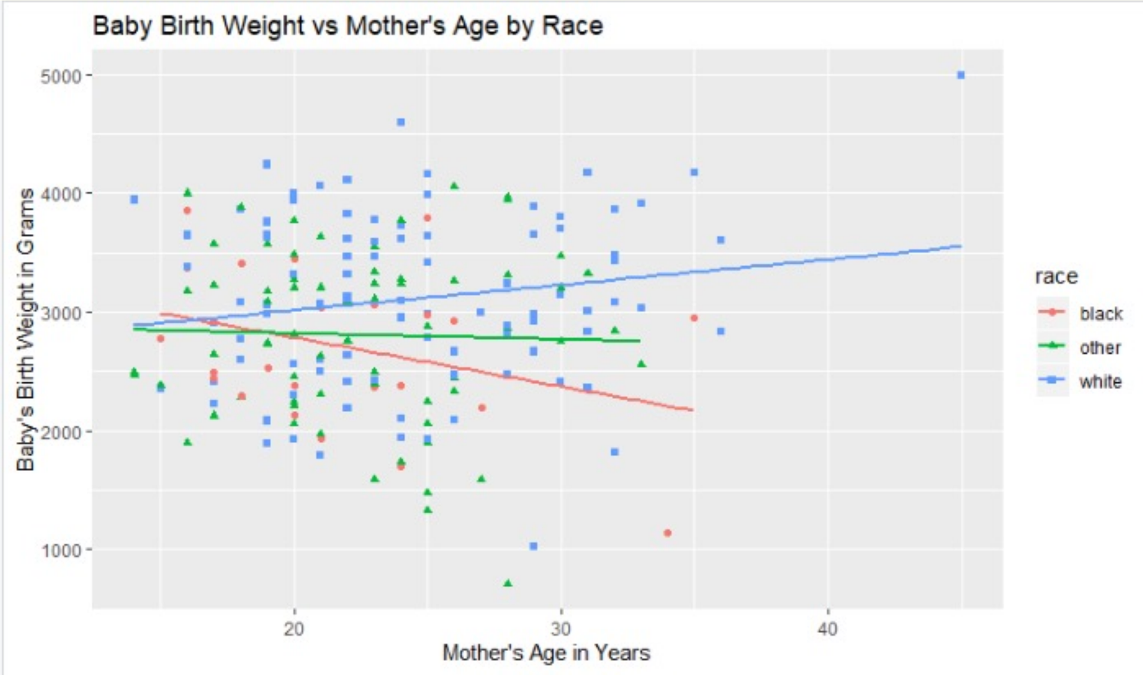
- We will use a data set on **birth-weight of babies** to show a situation where there IS interaction between a categorical variable (Race) and a continuous variable (Mother's Age) in **determining the birth-weight** of babies.
- The data set is available in Canvas under Lecture 4 Part D as "birth-weight.csv"
- The data were collected at Baystate Medical Center, Springfield, Mass during 1986. (**Source**: Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression.* New York: Wiley)

- **low** – baby's indicator of birth weight less than 2.5 kg.
- **age** - mother's age in years.
- **lwt** - mother's weight in pounds at last menstrual period.
- **race** - mother's race (1 = white, 2 = black, 3 = other).
- **smoke** – mother's smoking status during pregnancy.
- **ptl** – mother's number of previous premature labours.
- **ht** – mother's history of hypertension.
- **ui** – mother's presence of uterine irritability.
- **ftv** – mother's number of physician visits during the first trimester.
- **bwt** – baby's birth weight in grams.

| low | age | lwt | race | smoke | ht | ui | ftv | ptl | bwt |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 19 | 182 | black | 0 | 0 | 1 | 0 | 0 | 2523 |
| 0 | 33 | 155 | other | 0 | 0 | 0 | 1 | 0 | 2551 |
| 0 | 20 | 105 | white | 1 | 0 | 0 | 1 | 0 | 2557 |
| 0 | 21 | 108 | white | 1 | 0 | 1 | 1 | 0 | 2594 |
| 0 | 18 | 107 | white | 1 | 0 | 1 | 0 | 0 | 2600 |
| 0 | 21 | 124 | other | 0 | 0 | 0 | 0 | 0 | 2622 |
| 0 | 22 | 118 | white | 0 | 0 | 0 | 1 | 0 | 2637 |

# Multiple Regression – Interaction between Categorical & Continuous Predictors (**Interact.R**)

- We plot the Interaction Between Race and Age in Determining Birth-weight (bwt)
- You will now see three regression lines of bwt vs age, one for each sub-group of Race (White, Black, Other).
- The baby birth weight decreases with mother's age for race = "black", stays the same for race = "other" and increases for race = "white"
- This is a classic example of interaction. Notice that the **slopes** of the relationship between bwt and Age, depends on the race of the mother. *However, the interaction may or may* ***not*** *be statistically significant.*
- We have to run the regression model with the interaction term to see if the interaction is significant.




```
#
library(ggplot2)
#
ggplot(df, aes(x=age, y=bwt, shape=race, color=race)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Baby Birth Weight vs Mother's Age by Race") +
  xlab("Mother's Age in Years") +
  ylab("Baby's Birth Weight in Grams")
#
```

# Multiple Regression – Interaction between Categorical & Continuous Predictors (**Interact.R**)

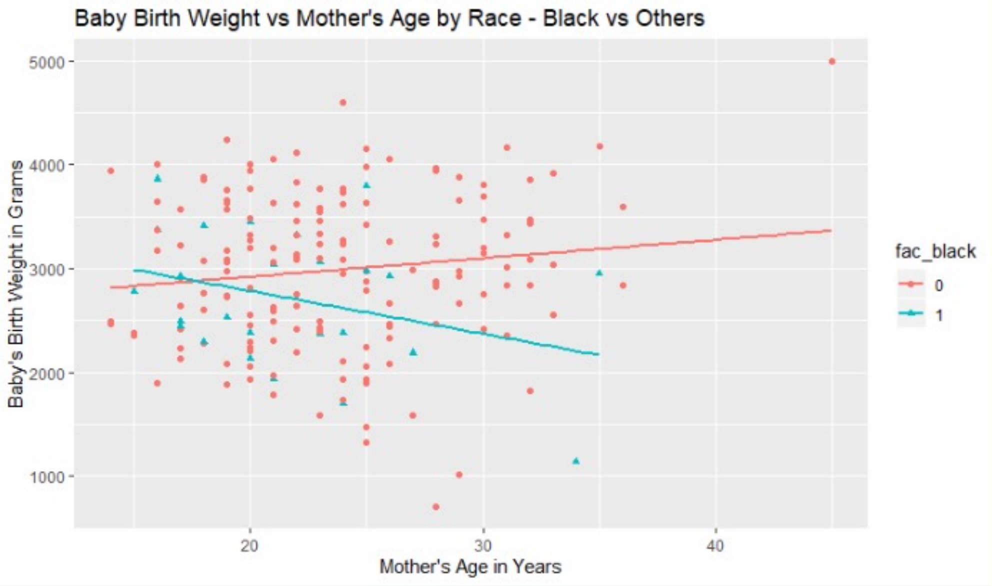
- To run the regression model, since Race is a categorical variable and we first generate dummy variables d_black (1 if Race=black, else 0), similarly White and Other.

- We will only focus on Black (vs White and Others) so we will only use that dummy variable and the interaction term.

e*d_black

```
> d_black <- ifelse(race == "black", 1, 0)
> #
> fac_black <- factor(d_black) # fac_black is only needed for ggplot -
> # the regression model uses d_black as dummy variable
> #
> #
> ggplot(df, aes(x=age, y=bwt, shape=fac_black, color=fac_black)) +
+   geom_point() +
+   geom_smooth(method = "lm", se = FALSE) +
+   ggtitle("Baby Birth Weight vs Mother's Age by Race - Black vs Others") +
+   xlab("Mother's Age in Years") +
+   ylab("Baby's Birth Weight in Grams")
> #
> # Regression Model with Interaction Term
> #
> mod1 <- lm(bwt ~ age+d_black+age*d_black)
> summary(mod1)
```

```
Call:
lm(formula = bwt ~ age + d_black + age * d_black)

Residuals:
     Min       1Q   Median       3Q      Max
-2351.28  -476.44    13.48   572.50  1627.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2563.09     257.43   9.956   <2e-16 ***
age            17.76      10.68   1.662   0.0982 .
d_black      1043.24     674.24   1.547   0.1235
age:d_black   -58.92      30.14  -1.955   0.0521 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 719.8 on 185 degrees of freedom
Multiple R-squared:  0.04066,   Adjusted R-squared:  0.0251
F-statistic: 2.614 on 3 and 185 DF,  p-value: 0.05265
```



# Multiple Regression – Interaction between Categorical & Continuous Predictors (**Interact.R**)

- For this example, let us assume that $\alpha = 0.10$. Based on this $\alpha$, the interaction term is significant.

- How do we interpret the interaction?

- First of all, it says that if we are interested in the effect of Mother's age on the birth-weight, *it depends on whether the mother's race is black or not*. Similarly, if we are interested in the effect of Mother's race on the birth-weight, *it depends on the mother's age*.

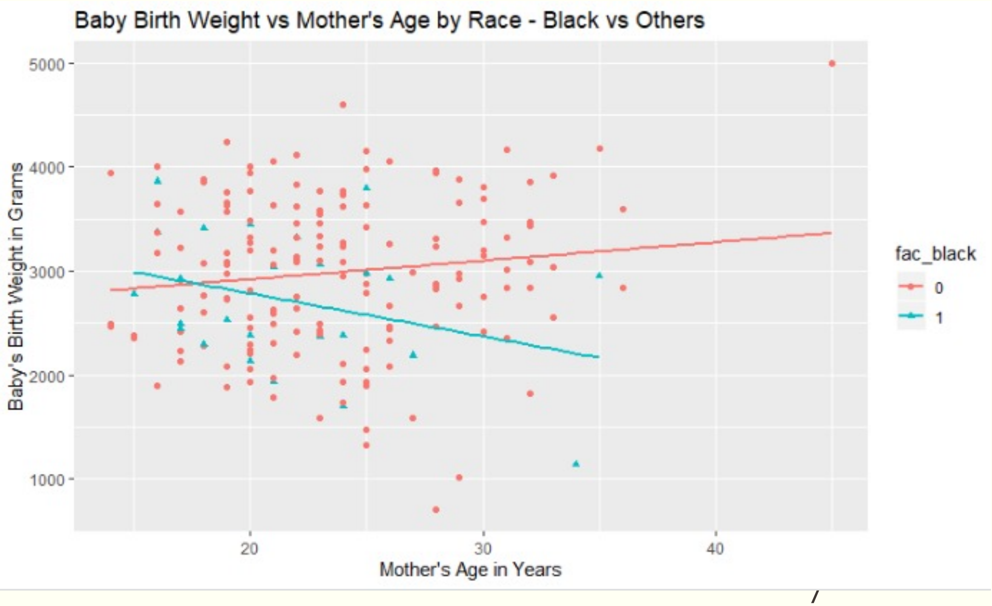
- Equations:
  - $\widehat{bwt} = 2563 + 17.76age$ (for Mother's Race NOT black)
  - $\widehat{bwt} = 3606 - 41.2age$ (for Mother's race = Black)

- That is for black mothers, the birth weight of the baby starts off higher (3606)and then decreases by -41.2 with each year's increase in Mother's age

- For Mother's who are not black, the birth weight of the baby starts off lower (2563) and then increases by 17.57 for each year's an increase in Mother's age

- In sum, rather than interpret main effects, even if they are significant, when interaction is significant, it is best to interpret the equations. You don't have to interpret the main effects such as age and d_black independently.

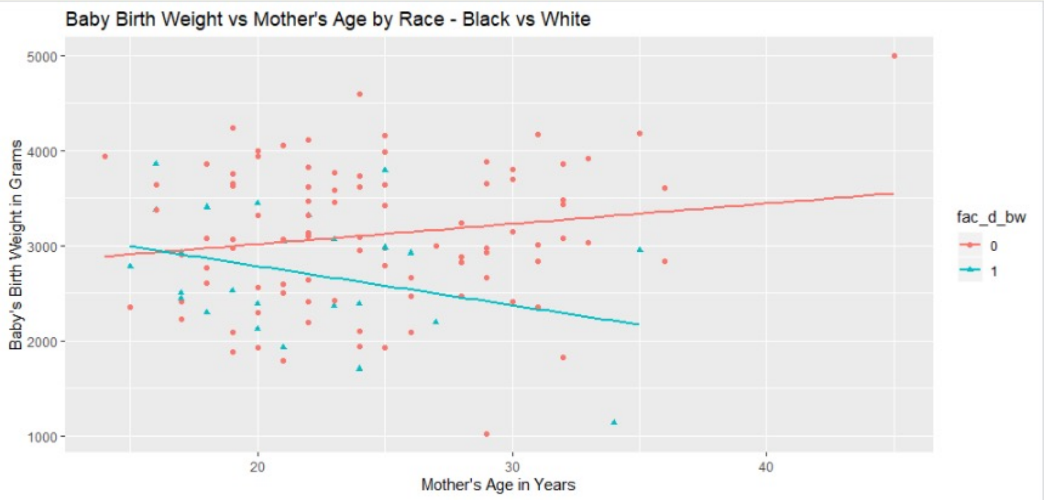
```
Call:
lm(formula = bwt ~ age + d_black + age * d_black)

Residuals:
     Min       1Q   Median       3Q      Max
-2351.28  -476.44    13.48   572.50  1627.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2563.09     257.43   9.956   <2e-16 ***
age            17.76      10.68   1.662   0.0982 .
d_black      1043.24     674.24   1.547   0.1235
age:d_black   -58.92      30.14  -1.955   0.0521 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 719.8 on 185 degrees of freedom
Multiple R-squared:  0.04066,   Adjusted R-squared:  0.0251
F-statistic: 2.614 on 3 and 185 DF,  p-value: 0.05265
```



# Multiple Regression – Interaction between Categorical & Continuous Predictors (**Interact.R**)

- We run the previous example comparing Black and White Mothers only.
- We notice that the interaction term is significant at $\alpha = 0.05$ since the p-value is 0.04.
- Equations:
  - $\widehat{bwt}$ = 2584.45 + 21.38age (for Mother's Race = White)
  - $\widehat{bwt}$ = 3606 - 41.2age (for Mother's race = Black)
- That is for black mothers, the birth weight of the baby starts off higher (3606)and then decreases by -41.2 with each year's increase in Mother's age
- For Mother's who are not black, the birth weight of the baby starts off lower (2584.45 and then increases by 21.38 for each year's an increase in Mother's age
- In sum, rather than interpret main effects, even if they are significant, when interaction is significant, it is best to interpret the equations. You don't have to interpret the main effects such as age and d_black independently.

```
> # Regression Model with Interaction Term - Black vs White
> #
> dfbw <- subset(df, race == "black" | race == "white", select=c(bwt, age, race))
> d_bw <- ifelse(dfbw$race == "black", 1, 0)
> fac_d_bw = factor(d_bw)
> #
> ggplot(dfbw, aes(x=age, y=bwt, shape=fac_d_bw, color=fac_d_bw)) +
+   geom_point() +
+   geom_smooth(method = "lm", se = FALSE) +
+   ggtitle("Baby Birth Weight vs Mother's Age by Race - Black vs White") +
+   xlab("Mother's Age in Years") +
+   ylab("Baby's Birth Weight in Grams")
> #
> #
> mod_bw <- lm(bwt ~ age+d_bw+age*d_bw, data=dfbw)
> summary(mod_bw)

Call:
lm(formula = bwt ~ age + d_bw + age * d_bw, data = dfbw)

Residuals:
     Min       1Q   Median       3Q      Max 
-2183.39  -473.57     2.99   553.60  1495.50 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2584.45     317.21   8.147 4.41e-13 ***
age            21.38      12.72   1.680   0.0955 .  
d_bw         1021.87     684.91   1.492   0.1384    
age:d_bw      -62.54      30.26  -2.067   0.0409 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 701.2 on 118 degrees of freedom
Multiple R-squared:  0.08676,	Adjusted R-squared:  0.06354 
F-statistic: 3.737 on 3 and 118 DF,  p-value: 0.01313
```



# INTERACTION BETWEEN CONTINUOUS PREDICTORS

Lecture 4D-2

# Multiple Regression – Interaction between Continuous Predictors (**Interact.R**)

- We will use a different data set on ice-cream consumptions to show a situation where there IS interaction between *two continuous variables* (price and family income) in **determining the per-capita consumption** of ice-cream.

- The data set is available in Canvas under Lecture 4 Part D as "icecream.csv".

- **Reference**: Koteswara Rao Kadiyala (1970) Testing for the independence of regression disturbances. *Econometrica*, 38, 97-117.

IC: Ice cream consumption in pints per capita
PRICE: Per pint price of ice cream in dollars
INC: Weekly family income in dollars
TEMP: Mean temperature in degrees F

```
> df1 <- read.table('icecream.csv',
+                   header = TRUE, sep = ',')
> print(df1)
      IC price income temp
1  0.386 0.270     78   41
2  0.374 0.282     79   56
3  0.393 0.277     81   63
4  0.425 0.280     80   68
5  0.406 0.272     76   69
6  0.344 0.262     78   65
7  0.327 0.275     82   61
8  0.288 0.267     79   47
9  0.269 0.265     76   32
10 0.256 0.277     79   24
11 0.286 0.282     82   28
12 0.298 0.270     85   26
13 0.329 0.272     86   32
14 0.318 0.287     83   40
15 0.381 0.277     84   55
16 0.381 0.287     82   63
17 0.470 0.280     80   72
18 0.443 0.277     78   72
19 0.386 0.277     84   67
20 0.342 0.277     86   60
21 0.319 0.292     85   44
22 0.307 0.287     87   40
23 0.284 0.277     94   32
24 0.326 0.285     92   27
25 0.309 0.282     95   28
26 0.359 0.265     96   33
27 0.376 0.265     94   41
28 0.416 0.265     96   52
29 0.437 0.268     91   64
30 0.548 0.260     90   71
>
```

# Multiple Regression – Interaction between Continuous Predictors (**Interact.R**)

- ***Just to illustrate*** the presence of interaction, I converted Income to a categorical variable such that it is hi_inc if Income is greater than 85,.

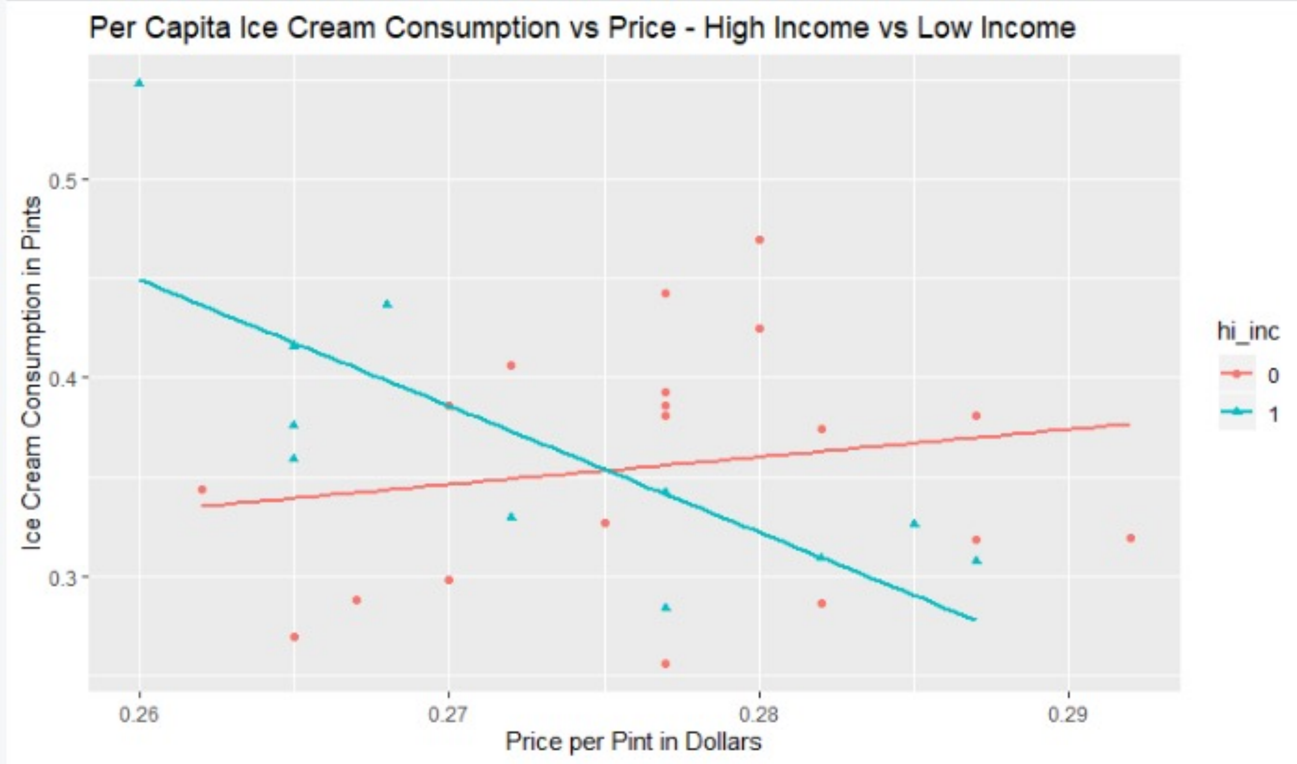
- The graph of Ice Cream Consumption (IC) vs Price is shown for the two income categories, along with the fit lines

- We can clearly see a difference in fit lines in both slope and intercept.
  - For hi_inc (higher income) low-priced ice-cream consumption is high at low prices and *consumption decreases as price increases.*
  - The *opposite is true for not hi_inc* (i.e., when the income is <= 85).

- This suggests an interaction between price and income in determining ice cream consumption.

- Note that income is a continuous variable. *We only converted it to a categorical variable to visually illustrate interaction with price.*

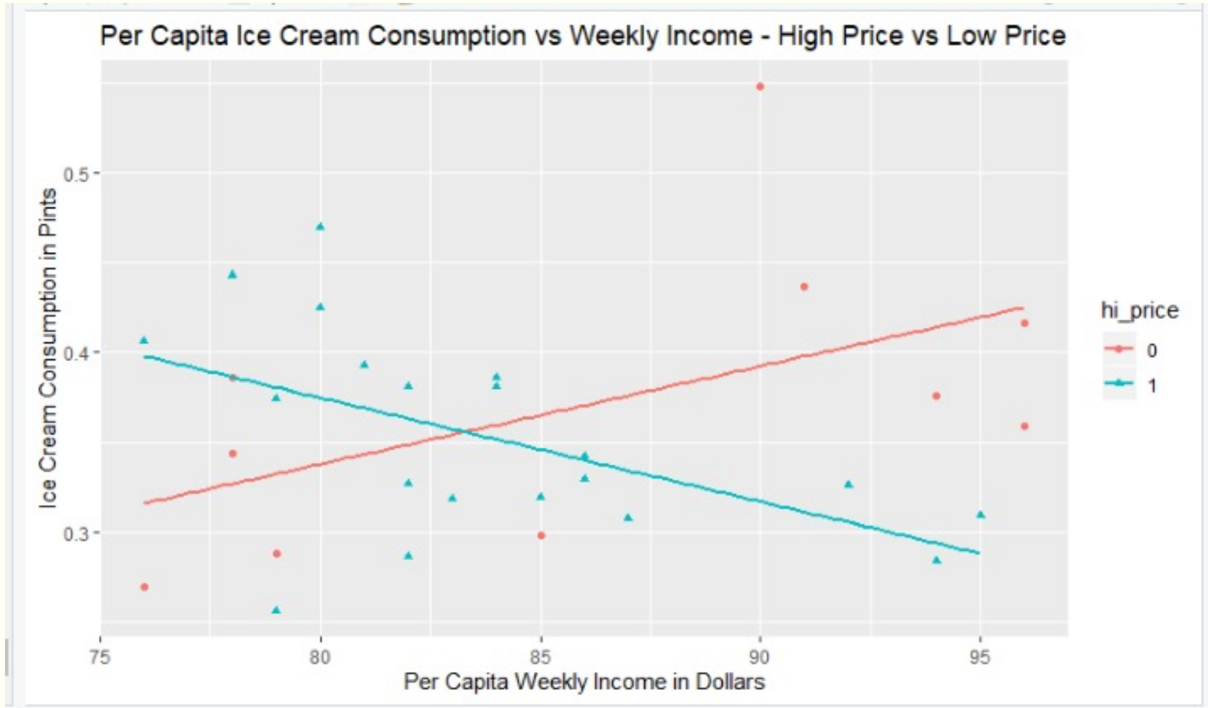
```
hi_inc <- factor(ifelse(df1$income >85, 1, 0))
ggplot(df1, aes(x=price, y=IC, shape=hi_inc, color=hi_inc)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Per Capita Ice Cream Consumption vs Price - High Income vs Low Income") +
  xlab("Price per Pint in Dollars") +
  ylab("Ice Cream Consumption in Pints")
```



# Multiple Regression – Interaction between Continuous Predictors (**Interact.R**)

- ***The same interaction*** between price and income in determining IC can also be shown by converting price to a categorical variable.
- I converted price to a categorical variable such that hi_price = price > 27.
- Again, we can see that consumer behavior for price > 27 (hi_price = 1) increases with income while for price <= 27 (hi_price = 0) it decreases with income.
- *This gives us the same information as before – Ice Cream Consumption depends on the values of both Price and Income.*

```
hi_price <- factor(ifelse(df1$price > 0.27,1,0))
ggplot(df1, aes(x=income, y=IC, shape=hi_price, color=hi_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Per Capita Ice Cream Consumption vs Weekly Income - High Price vs Low Price") +
  xlab("Per Capita Weekly Income in Dollars") +
  ylab("Ice Cream Consumption in Pints")
```



# Multiple Regression – Interaction between Continuous Predictors (**Interact.R**)

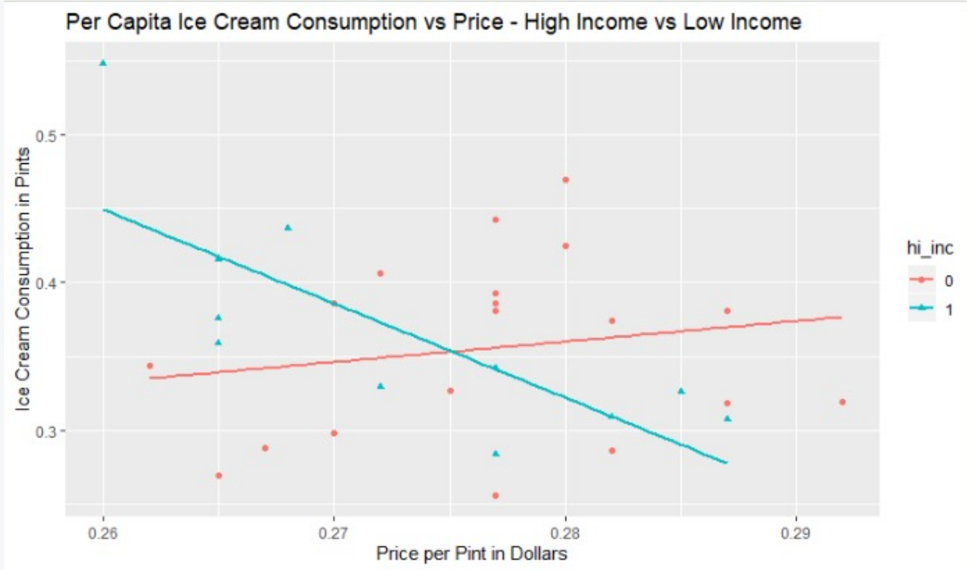
- Our Model will be:
  - $\widehat{ic} = \widehat{\alpha} + \widehat{\beta}_{price}price + \widehat{\beta}_{income}income + \widehat{\beta}_{incprice}income*price$
  - $\widehat{ic} = -15.3 + 57.58\ price + 0.189\ income - 0.696\ income*price$
- We notice that the predictors explain 30.73% of the variability in ice cream consumption
- The overall model fit given by the ANOVA F-test is significant at **α = 0.05** indicating that the model does explain a significant proportion of the variability in ice cream consumption, relative to the error term

```
> # Regression model with interaction
> #
> mod2 <- lm(IC ~ price+income+price*income, data = df1)
> summary(mod2)

Call:
lm(formula = IC ~ price + income + price * income, data = df1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.116865 -0.038381 -0.007629  0.033801  0.127481 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -15.3044     5.4209  -2.823  0.00900 **
price         57.5837    19.9303   2.889  0.00769 **
income         0.1893     0.0631   3.000  0.00589 **
price:income  -0.6958     0.2321  -2.997  0.00592 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05783 on 26 degrees of freedom
Multiple R-squared:  0.3072,    Adjusted R-squared:  0.2273 
F-statistic: 3.843 on 3 and 26 DF,  p-value: 0.02113

> print(anova(mod2))
Analysis of Variance Table

Response: IC
             Df   Sum Sq   Mean Sq F value   Pr(>F)   
price         1 0.008459 0.0084589  2.5290 0.123855   
income        1 0.000051 0.0000510  0.0152 0.902702   
price:income  1 0.030051 0.0300511  8.9847 0.005922 **
Residuals    26 0.086962 0.0033447                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

# Multiple Regression – Interaction between Continuous Predictors (**Interact.R**)

- The first thing we notice is that the interaction term is significant at $\alpha$ = 0.05.
- This means that price and income *interact in determining* the consumption of ice cream (IC).
- $\widehat{ic}$ = -15.3 + 57.58 price + 0.189 income – 0.696 income*price
- This means that the consumption of ice cream (IC), *depends on the both the price and the income of the buyer.*
- Even though the main effects of price and income are significant at $\alpha$ = 0.05 (p-value for price = 0.0090 and p-value for income is 0.0077), we cannot independently assess their effects on price and therefore we do not interpret them independently.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -15.3044     5.4209  -2.823  0.00900 **
price         57.5837    19.9303   2.889  0.00769 **
income         0.1893     0.0631   3.000  0.00589 **
price:income  -0.6958     0.2321  -2.997  0.00592 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05783 on 26 degrees of freedom
Multiple R-squared:  0.3072,    Adjusted R-squared:  0.2273
F-statistic: 3.843 on 3 and 26 DF,  p-value: 0.02113
```



# Predictions of Ice Cream Consumption

- To show you the interesting consequences of the interaction effect we compare the ice cream consumption of a low income consumer vs high income consumers for high priced ice cream.
- It clearly shows that lower income consumers actually consume more of high priced ice cream than high income consumers.
- The converse is also true.

```
> #
> # Prediction of Ice Cream consumption by a low income consumer for high priced ice-cream
> #
> df_a <- data.frame(price=0.292, income=76)
> print(predict(mod2, df_a))
       1
0.453735
> # Prediction of Ice Cream consumption by a High income consumer for high priced ice-cream
> #
> df_b <- data.frame(price=0.292, income=96)
> print(predict(mod2, df_b))
        1
0.1757535
```

# Multiple Regression – Interaction between Continuous Predictors

- In interaction models it is always a good idea to check for Multicollinearity.
- In this case both the predictors (income and price) appear in the interaction term and we expect the interaction term to be highly correlated with both predictors.
- Remember, we want the variance inflation factors (VIFs) to be close to 1 and VIFs between 5 and 10 indicate high multicollinearity
- Our collinearity statistics show VIFs off the charts! This indicates excessive *Multicollinearity*.
- *If we are going to use the model only to generate predictions*, we do not need to worry about the multicollinearity.
- *But, if we want to interpret the slope coefficients* (i.e., the effect of price and income on IC), we need to do something about it.

```
> library(olsrr)

Attaching package: ‘olsrr’

The following object is masked from ‘package:MASS’:

    cement

The following object is masked from ‘package:datasets’:

    rivers

> ols_vif_tol(mod2)
# A tibble: 3 x 3
  Variables    Tolerance   VIF
  <chr>            <dbl> <dbl>
1 price         0.00417   240.
2 income        0.000743 1347.
3 price:income  0.000683 1464.
>
```

# Multiple Regression – Interaction between Continuous Predictors

- **Handling Multicollinearity in Interaction Models**:

  - One option to handle multicollinearity in interaction between continues predictors that is frequently exercised for continuous predictors is to "center" them. i.e., subtract the mean from each observation, and use the "centered variable" as a predictor.

  - We create centered predictors c_price and c_income as:
    - c_price = price – mean(price) and similarly c_income. = income – mean(income)
    - The interaction term now is cpr_cinc = c_price*c_income

  - We show the correlation matrix between the predictors and interaction term, before and after centering.

  - The correlation matrix for cprice, cincome shows reduced correlations among the centered predictors and the interaction term cpr_cinc.

```
> price <- df1$price
> income <- df1$income
> pr_inc <- price*income
> A <- cbind(price, income, pr_inc)
> print (cor(A))
            price     income    pr_inc
price   1.0000000 -0.107479 0.3007737
income -0.1074790  1.000000 0.9154880
pr_inc  0.3007737  0.915488 1.0000000
```

```
> c_price <- price - mean(price)
> c_income <- income - mean(income)
> cpr_cinc <- c_price*c_income
> #
> # Correlation Matrix of Centered Variables
> B <- cbind(c_price, c_income, cpr_cinc)
> #
> print (cor(B))
            c_price  c_income   cpr_cinc
c_price   1.0000000 -0.107479  0.2156068
c_income -0.1074790  1.000000 -0.4437530
cpr_cinc  0.2156068 -0.443753  1.0000000
```

# Multiple Regression – Interaction between Continuous Predictors

- We run our model with the centered predictors and their interaction.
- The model fit and the $R^2$ of the model will be unchanged from the non-centered case.
- The interaction term is significant at **α = 0.05**, *but not the centered main effects*
- **The VIFS are in line** (close to 1) showing that there is no Multicollinearity.
- The intercept will be different (-15.304 without centering) and 0.356 for the centered variables model.

```
> mod3 <- lm(IC ~ c_price+c_income+cpr_cinc, data = df1)
> summary(mod3)

Call:
lm(formula = IC ~ c_price + c_income + cpr_cinc, data = df1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.116865 -0.038381 -0.007629  0.033801  0.127481 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.355666   0.010633  33.448  < 2e-16 ***
c_price     -1.285239   1.318446  -0.975  0.33864    
c_income    -0.002278   0.001919  -1.187  0.24585    
cpr_cinc    -0.695851   0.232148  -2.997  0.00592 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05783 on 26 degrees of freedom
Multiple R-squared:  0.3072,	Adjusted R-squared:  0.2273 
F-statistic: 3.843 on 3 and 26 DF,  p-value: 0.02113
```

```
> ols_vif_tol(mod3)
# A tibble: 3 x 3
  Variables Tolerance   VIF
  <chr>         <dbl> <dbl>
1 c_price       0.953  1.05
2 c_income      0.803  1.25
3 cpr_cinc      0.775  1.29
> 
```

# Multiple Regression – Interaction between Continuous Predictors

- Our Original (non-centered) Model will be:
  - $\widehat{ic} = -15.304 + 57.584price + 0.189income - 0.696income*price$
- Our Centered Model will be:
  - $\widehat{ic} = 0.356 - 1.285c\_price - 0.002c\_income - 0.696c\_income*c\_price$
- Both models will give the same predictions once you substitute
  - c_price = price – mean(price) and c_income = income – mean(income).
  - c_price = price – 0.2753 and c_income = income – mean(84.6).
- The slope of the interaction term, or its t-statistic and significance, will not change for the centered model relative to the original non-centered model, but the main effect slopes will change, as well as the t-statistics.

# Final Model

- Remember that if the model is only used for prediction, Multicollinearity is not an issue. It is only interpretation of the coefficients that can be problematic. But the centered model also makes it more difficult to interpret coefficients.

- If we are not interested in predictions but in having a model where the predictors are significant, we may choose a cantered model with only the significant interaction term.

- The $R^2$ falls from 0.3072 to 0.2451 but the Adjusted $R^2$ falls from 0.2273 to 0.2182 which is not a substantial reduction.

- So the model:
  - $\widehat{ic} = 0.3561 - 0.6197(\text{income} - 0.2753)*(\text{price} - 84.6)$
  - $\widehat{ic} = 0.3561 - 0.6197(\text{income}*\text{price}) + 0.1760\text{price} + 52.4266\text{income}$

- It is clear that the interaction term from the centered variables contains main effects and is not an interaction only model when centering is removed.

- So we can use the centered model, or go back to the original model and note that in the presence of significant interaction, we have to be cautious in interpreting the main effects.

- The most important outcome of our analysis is that:
  - "Higher income" and "Middle Income" buyers consume less as the price of ice cream increases
  - "Lower income" buyers consume *more* as the price of ice cream increases

Centered model

```
> # Final centered model with only significant interaction term
> #
> mod4 <- lm(IC ~ cpr_cinc, data = df1)
> summary(mod4)

Call:
lm(formula = IC ~ cpr_cinc, data = df1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.105978 -0.037691 -0.008967  0.040021  0.140722 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.35608    0.01068  33.344  < 2e-16 ***
cpr_cinc    -0.61970    0.20551  -3.015  0.00541 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05817 on 28 degrees of freedom
Multiple R-squared:  0.2451,	Adjusted R-squared:  0.2182 
F-statistic: 9.093 on 1 and 28 DF,  p-value: 0.005407
```

Original model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -15.3044     5.4209  -2.823  0.00900 **
price         57.5837    19.9303   2.889  0.00769 **
income         0.1893     0.0631   3.000  0.00589 **
price:income  -0.6958     0.2321  -2.997  0.00592 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05783 on 26 degrees of freedom
Multiple R-squared:  0.3072,	Adjusted R-squared:  0.2273 
F-statistic: 3.843 on 3 and 26 DF,  p-value: 0.02113
```

# Summary of Multiple Regression with Interaction Terms

- Interaction means that the effect of each predictor in the interaction term on the dependent variable, *depends upon the level/value of the other predictor* in the interaction term.

- The tests for $H_0$: ($\beta_{price}$) = **0** and $H_0$: ($\beta_{income}$) = **0** are called tests of *main effects (as opposed to interaction)*

- The main effects have separate effects on the dependent variable, controlling for other main effects and interactions in the model. That is a defined proportion of the variability in the dependent variable is attributable to each main effect, *regardless of the levels/values of the other main effect.*

- However, if interaction is significant, interpreting the main effects becomes difficult.

- Consider: $\widehat{ic}$ = -15.3 + 57.58 price + 0.189 income – 0.696 income*price
  - We can re-write this as: $\widehat{ic}$ = -15.3 + (57.58 – 0.696income)*price + 0.189 income
  - That is the slope of price = (57.58 – 0.696income)
  - The interpretation of slope of price is the change in predicted IC ($\widehat{ic}$) for a unit change in price.
  - Clearly this depends on what income is. This is the meaning of interaction itself.
  - Thus, interpreting the main effect of price is not easy because it is not independent of income.
  - This is why it is better to focus just on predictions of IC.

- If the interaction term is *not significant*, we can look at the significance of the individual main effects of the predictors in the interaction term and we can interpret the beta coefficients of the main effects.

- We have only focused on pairwise interactions (interactions between two predictors) but more complicated interactions among more than 2 predictors is possible.

- All these issues are relevant whether the predictors involved in the interaction are categorical and/or continuous.