# 05-MLR-Part2

Objectives:

1. Review MLR with continuous and categorical variable
2. Fit a MLR with interaction term with continuous variables
3. Fit a MLR with interaction terms with continuous and categorical variable
4. Fit a MLR with 3-way interaction term and 2-way interaction terms

## Load packages

```
library(tidyverse) # data manipulation
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)   # plotting
```

## Load data

```
data <- read.csv("train.csv")
```

## 1. Review MLR with continuous and categorical variable

Refer to **04-MLR.Rmd** (**01-EDA-02.Rmd** for now) or proceed to 2.

## 2. Fit a MLR with interaction term with continuous variables

Recall that in fit2, we fit `TotalBsmtSF` and `LotArea`. What if we have reason to suspect that the effect of `TotalBsmtSF` on `SalePrice`, and vice versa?

fit2 is also referred to as an additive model since the effect is additive. It is also known as the parallel- or same-slopes model.

```
fit2 <- lm(SalePrice ~ TotalBsmtSF + LotArea, data = data)
summary(fit2)
```

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF + LotArea, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -603042  -38254  -13652   32500  417866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.967e+04  4.308e+03  13.850  < 2e-16 ***
## TotalBsmtSF 1.058e+02  3.844e+00  27.534  < 2e-16 ***
## LotArea     8.865e-01  1.690e-01   5.247 1.78e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62190 on 1457 degrees of freedom
## Multiple R-squared:  0.388,  Adjusted R-squared:  0.3872
## F-statistic: 461.9 on 2 and 1457 DF,  p-value: < 2.2e-16
```

fit8 extends fit2 to include an interaction term. fit8 is referred to as the non-additive model. It is also known as different-slopes model.

```
fit8 <- lm(SalePrice ~ TotalBsmtSF + LotArea + TotalBsmtSF:LotArea, data = data)
summary(fit8)
```

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF + LotArea + TotalBsmtSF:LotArea,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197729  -38015  -12735   35940  423839
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.440e+04  5.886e+03   2.446   0.0146 *
## TotalBsmtSF          1.375e+02  4.713e+00  29.173   <2e-16 ***
## LotArea              4.659e+00  3.842e-01  12.126   <2e-16 ***
## TotalBsmtSF:LotArea -2.273e-03  2.097e-04 -10.837   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59840 on 1456 degrees of freedom
## Multiple R-squared:  0.4337, Adjusted R-squared:  0.4326
## F-statistic: 371.7 on 3 and 1456 DF,  p-value: < 2.2e-16
```

The p-val associated with `TotalBsmtSF:LotArea` is significant (`p-val = <2e-16`). Thus, we reject the null hypothesis Ho that $\beta_3 = 0$ and conclude that the effect of `TotalBsmtSF` on `SalePrice` depends on `LotArea`, and vice versa.

Note again that the fit is poor (`R-squared: 0.06111`, `Adjusted R-squared: 0.05725`).

## 3. Fit a MLR with interaction terms with continuous and categorical variable

Recall that fit7 fits a continuous variable and a categorical variable. Can you guess how many parameters in total if we were to fit an interaction term?

Notice that in **01-EDA-02.Rmd** we briefly cleaned the data and rename the data set as data_sub already. We need to do additional cleaning later but let's ignore it for now.

```
data$MSZoning %>% str()
```

```
##  chr [1:1460] "RL" "RL" "RL" "RL" "RL" "RL" "RL" "RL" "RM" "RL" "RL" "RL" ...
```

```
data$MSZoning <- factor(data$MSZoning)
```

```
data$MSZoning %>% str()
```

```
##  Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
```

```
levels(data$MSZoning)
```

```
## [1] "C (all)" "FV"      "RH"      "RL"      "RM"
```

fit7 has a total of parameters: 1 ($\beta_0$) for the intercept + 1 ($\beta_1$) for `TotalBsmtSF` + (5 - 1) ($\beta_2$, $\beta_3$, $\beta_4$, $\beta_5$) for each level/group of `MSZoning` excluding the reference group.

```
fit7 <- lm(SalePrice ~ TotalBsmtSF + MSZoning, data = data)
summary(fit7)
```

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF + MSZoning, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -550854  -39104  -10763   28846  425750
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5665.184  19484.791  -0.291   0.7713
## TotalBsmtSF    104.093      3.776  27.565  < 2e-16 ***
## MSZoningFV  116521.562  20712.169   5.626 2.21e-08 ***
## MSZoningRH   51060.646  24560.860   2.079   0.0378 *
## MSZoningRL   80511.616  19393.828   4.151 3.50e-05 ***
## MSZoningRM   48866.207  19703.595   2.480   0.0132 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60930 on 1454 degrees of freedom
## Multiple R-squared:  0.4139, Adjusted R-squared:  0.4118
## F-statistic: 205.3 on 5 and 1454 DF,  p-value: < 2.2e-16
```

fit9 extends fit 7 to include an interaction term.

```
fit9 <- lm(SalePrice ~ TotalBsmtSF + MSZoning + TotalBsmtSF:MSZoning, data = data)
summary(fit9)
```

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF + MSZoning + TotalBsmtSF:MSZoning,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -556292  -38851  -10972   29084  424304
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -9866.300  72143.129  -0.137    0.891
## TotalBsmtSF                109.546     90.238   1.214    0.225
## MSZoningFV              121383.250  76241.524   1.592    0.112
## MSZoningRH              102505.340  83355.064   1.230    0.219
## MSZoningRL               83497.568  72302.030   1.155    0.248
## MSZoningRM               60535.905  73228.941   0.827    0.409
## TotalBsmtSF:MSZoningFV      -6.120     93.294  -0.066    0.948
## TotalBsmtSF:MSZoningRH     -62.528    101.727  -0.615    0.539
## TotalBsmtSF:MSZoningRL      -4.364     90.325  -0.048    0.961
## TotalBsmtSF:MSZoningRM     -14.807     91.453  -0.162    0.871
##
## Residual standard error: 60970 on 1450 degrees of freedom
## Multiple R-squared:  0.4146, Adjusted R-squared:  0.411
## F-statistic: 114.1 on 9 and 1450 DF,  p-value: < 2.2e-16
```

## 4. Fit a MLR with 3-way interaction term and 2-way interaction terms

```
data %>% glimpse
```

```
## Rows: 1,460
## Columns: 81
## $ Id            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ MSSubClass    <int> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 60, 20, 20,~
## $ MSZoning      <fct> RL, RL, RL, RL, RL, RL, RL, RL, RM, RL, RL, RL, RL, RL, ~
## $ LotFrontage   <int> 65, 80, 68, 60, 84, 85, 75, NA, 51, 50, 70, 85, NA, 91, ~
## $ LotArea       <int> 8450, 9600, 11250, 9550, 14260, 14115, 10084, 10382, 612~
## $ Street        <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", ~
## $ Alley         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ LotShape      <chr> "Reg", "Reg", "IR1", "IR1", "IR1", "IR1", "Reg", "IR1", ~
## $ LandContour   <chr> "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", ~
## $ Utilities     <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub", "AllPu~
## $ LotConfig     <chr> "Inside", "FR2", "Inside", "Corner", "FR2", "Inside", "I~
## $ LandSlope     <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", ~
## $ Neighborhood  <chr> "CollgCr", "Veenker", "CollgCr", "Crawfor", "NoRidge", "~
## $ Condition1    <chr> "Norm", "Feedr", "Norm", "Norm", "Norm", "Norm", "Norm",~
## $ Condition2    <chr> "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", ~
## $ BldgType      <chr> "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", ~
## $ HouseStyle    <chr> "2Story", "1Story", "2Story", "2Story", "2Story", "1.5Fi~
## $ OverallQual   <int> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7, 6, 4, 5,~
## $ OverallCond   <int> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, 7, 5, 5,~
## $ YearBuilt     <int> 2003, 1976, 2001, 1915, 2000, 1993, 2004, 1973, 1931, 19~
## $ YearRemodAdd  <int> 2003, 1976, 2002, 1970, 2000, 1995, 2005, 1973, 1950, 19~
## $ RoofStyle     <chr> "Gable", "Gable", "Gable", "Gable", "Gable", "Gable", "G~
## $ RoofMatl      <chr> "CompShg", "CompShg", "CompShg", "CompShg", "CompShg", "~
## $ Exterior1st   <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Sdng", "VinylSd", "~
## $ Exterior2nd   <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Shng", "VinylSd", "~
## $ MasVnrType    <chr> "BrkFace", "None", "BrkFace", "None", "BrkFace", "None",~
## $ MasVnrArea    <int> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286, 0, 306, ~
## $ ExterQual     <chr> "Gd", "TA", "Gd", "TA", "Gd", "TA", "Gd", "TA", "TA", "T~
## $ ExterCond     <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "T~
## $ Foundation    <chr> "PConc", "CBlock", "PConc", "BrkTil", "PConc", "Wood", "~
## $ BsmtQual      <chr> "Gd", "Gd", "Gd", "TA", "Gd", "Gd", "Ex", "Gd", "TA", "T~
## $ BsmtCond      <chr> "TA", "TA", "TA", "Gd", "TA", "TA", "TA", "TA", "TA", "T~
## $ BsmtExposure  <chr> "No", "Gd", "Mn", "No", "Av", "No", "Av", "Mn", "No", "N~
## $ BsmtFinType1  <chr> "GLQ", "ALQ", "GLQ", "ALQ", "GLQ", "GLQ", "GLQ", "ALQ", ~
## $ BsmtFinSF1    <int> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851, 906, 99~
## $ BsmtFinType2  <chr> "Unf", "Unf", "Unf", "Unf", "Unf", "Unf", "Unf", "BLQ", ~
## $ BsmtFinSF2    <int> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ BsmtUnfSF     <int> 150, 284, 434, 540, 490, 64, 317, 216, 952, 140, 134, 17~
## $ TotalBsmtSF   <int> 856, 1262, 920, 756, 1145, 796, 1686, 1107, 952, 991, 10~
## $ Heating       <chr> "GasA", "GasA", "GasA", "GasA", "GasA", "GasA", "GasA", ~
## $ HeatingQC     <chr> "Ex", "Ex", "Ex", "Gd", "Ex", "Ex", "Ex", "Ex", "Gd", "E~
## $ CentralAir    <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "~
## $ Electrical    <chr> "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "S~
## $ X1stFlrSF     <int> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022, 1077, ~
## $ X2ndFlrSF     <int> 854, 0, 866, 756, 1053, 566, 0, 983, 752, 0, 0, 1142, 0,~
## $ LowQualFinSF  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ GrLivArea     <int> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090, 1774, 10~
## $ BsmtFullBath  <int> 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1,~
## $ BsmtHalfBath  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```
## $ FullBath      <int> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, 1, 2, 1,~
## $ HalfBath      <int> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,~
## $ BedroomAbvGr  <int> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, 2, 2, 3,~
## $ KitchenAbvGr  <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1,~
## $ KitchenQual   <chr> "Gd", "TA", "Gd", "Gd", "Gd", "TA", "Gd", "TA", "TA", "T~
## $ TotRmsAbvGrd  <int> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5, 5, 6, 6~
## $ Functional    <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", ~
## $ Fireplaces    <int> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, 1, 0, 0,~
## $ FireplaceQu   <chr> NA, "TA", "TA", "Gd", "TA", NA, "Gd", "TA", "TA", "TA", ~
## $ GarageType    <chr> "Attchd", "Attchd", "Attchd", "Detchd", "Attchd", "Attch~
## $ GarageYrBlt   <int> 2003, 1976, 2001, 1998, 2000, 1993, 2004, 1973, 1931, 19~
## $ GarageFinish  <chr> "RFn", "RFn", "RFn", "Unf", "RFn", "Unf", "RFn", "RFn", ~
## $ GarageCars    <int> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2, 2,~
## $ GarageArea    <int> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205, 384, 7~
## $ GarageQual    <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "Fa", "G~
## $ GarageCond    <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "T~
## $ PavedDrive    <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "~
## $ WoodDeckSF    <int> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, 140, 160~
## $ OpenPorchSF   <int> 61, 0, 42, 35, 84, 30, 57, 204, 0, 4, 0, 21, 0, 33, 213,~
## $ EnclosedPorch <int> 0, 0, 0, 272, 0, 0, 0, 228, 205, 0, 0, 0, 0, 0, 176, 0, ~
## $ X3SsnPorch    <int> 0, 0, 0, 0, 0, 320, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ ScreenPorch   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 176, 0, 0, 0, 0, ~
## $ PoolArea      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ PoolQC        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Fence         <chr> NA, NA, NA, NA, NA, "MnPrv", NA, NA, NA, NA, NA, NA, NA,~
## $ MiscFeature   <chr> NA, NA, NA, NA, NA, "Shed", NA, "Shed", NA, NA, NA, NA, ~
## $ MiscVal       <int> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0, 0, 700,~
## $ MoSold        <int> 2, 5, 9, 2, 12, 10, 8, 11, 4, 1, 2, 7, 9, 8, 5, 7, 3, 10~
## $ YrSold        <int> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, 2008, 20~
## $ SaleType      <chr> "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "W~
## $ SaleCondition <chr> "Normal", "Normal", "Normal", "Abnorml", "Normal", "Norm~
## $ SalePrice     <int> 208500, 181500, 223500, 140000, 250000, 143000, 307000, ~
```

Instead of using :, we use * to shorten the line of code. Check how many parameters fit10 has?

```
fit10 <- lm(SalePrice ~ TotalBsmtSF * LotArea * GarageArea, data = data)
summary(fit10)
```

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF * LotArea * GarageArea,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -240749  -28131   -4819   25965  375846
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    8.678e+04  1.111e+04   7.812 1.08e-14 ***
## TotalBsmtSF                    8.863e+00  1.037e+01   0.855    0.393
## LotArea                        5.789e-02  8.971e-01   0.065    0.949
## GarageArea                    -3.201e+01  2.090e+01  -1.532    0.126
```

```
## TotalBsmtSF:LotArea             7.565e-04  4.921e-04   1.537    0.124
## TotalBsmtSF:GarageArea          1.604e-01  1.714e-02   9.363  < 2e-16 ***
## LotArea:GarageArea              5.342e-03  1.281e-03   4.170 3.22e-05 ***
## TotalBsmtSF:LotArea:GarageArea -3.783e-06  4.531e-07  -8.350  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51540 on 1452 degrees of freedom
## Multiple R-squared:  0.5812, Adjusted R-squared:  0.5792
## F-statistic: 287.8 on 7 and 1452 DF,  p-value: < 2.2e-16
```

Here we only have 3 predictor variables. Now imagine if we fit a model with 81 predictor variables...