TOPIC2: Multiple Linear Regression - Interaction Effects and Second-Order Models

An Interaction Model with Quantitative Predictors

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

According to this model, if we increase X_1 by one unit, then Y will increase by an average of β_1 units. Notice that the presence of X_2 does not alter this statement-that is, regardless of the value of X_2, a one-unit increase in X_1 will lead to a β_1-unit increase in Y. The above equation is also known as additive model, investigating only the main effects of predictors. It assumes that the relationship between a given predictor variable and the response is independent of the other predictor variable.

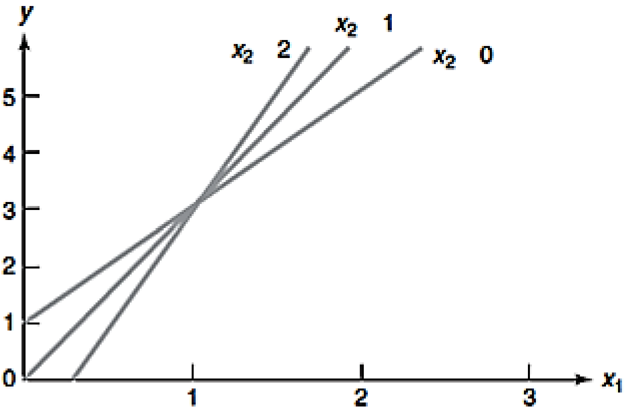
If the relationship between $E(Y)$ and X_1 depends on the values of the remaining X's held fixed, then the first-order model is not appropriate for predicting Y. Interaction occurs whenever the effect of an independent variable on a dependent variable is not constant over all of the values of the other independent variables. In this case, we need another model that will take into account this dependence. Such a model includes the cross products of two or more X's. Hence, the interaction model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

For example, suppose that the mean value $E(Y)$ of a response Y is related to two quantitative independent variables, X_1 and X_2, by the model

$$E(Y) = 1 + 2X_1 - X_2 + X_1 X_2$$

A graph of the relationship between $E(Y)$ and X_1 for $X_2 = 0$, 1, and 2 is displayed in Figure 1.



Note that the graph shows three nonparallel straight lines. You can verify that the slopes of the lines differ by substituting each of the values $X_2 = 0$, 1, and 2 into the equation.

For $X_2 = 0$:

$E(Y) = 1 + 2X_1 - (0) + X_1(0) = 1 + 2X_1(slope = 2)$

For $X_2 = 1$:

$E(Y) = 1 + 2X_1 - (1) + X_1(1) = 3X_1(slope = 3)$

For $X_2 = 2$:

$E(Y) = 1 + 2X_1 - (2) + X_1(2) = -1 + 4X_1(slope = 4)$

Note that the slope of each line is represented by slope=$\beta_1 + \beta_3 x_2 = 2 + x_2$. Thus, the effect on $E(Y)$ of a change in X_1 (i.e., the slope) now depends on the value of X_2. When this situation occurs, we say that X_1 and X_2 interact. Otherwise, the graph for 3 lines would be parallel. The cross-product term, $X_1 X_2$, is called an interaction term, and the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$ is called **an interaction model with two quantitative variables.**

Testing for Interaction in Multiple Regression

For testing an interaction term in regression model, we use the Individual Coefficients Test (t-test) method.

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0 \text{ (i=1,2,...,p)}$$

$$t_{cal} = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \text{ which has df=n-p degree of freedom}$$

where p is the total number of independent variables (including interaction terms).

Considering our advertising example, let's test the interation term.

```
Advertising=read.table("Advertising.txt",header = TRUE,  sep ="\t")
interacmodel<-lm(sale~tv+radio+tv:radio, data=Advertising)
summary(interacmodel)
```

```
##
## Call:
## lm(formula = sale ~ tv + radio + tv:radio, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
## radio       2.886e-02  8.905e-03   3.241   0.0014 **
## tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
#option2
interacmodel1<-lm(sale~tv*radio, data=Advertising)
summary(interacmodel1)
```

```
##
## Call:
## lm(formula = sale ~ tv * radio, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
## radio       2.886e-02  8.905e-03   3.241   0.0014 **
## tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
#option3
interacmodel2<-lm(sale~(tv + radio)^2, data=Advertising)
summary(interacmodel2)
```

```
##
```

```
## Call:
## lm(formula = sale ~ (tv + radio)^2, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
## radio       2.886e-02  8.905e-03   3.241   0.0014 **
## tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

As you can see from the output, t_{cal}= 20.727 with the p-value< 0.0001, indicating that we should clearly reject the null hypothesis which means that we should definetely add the interaction term to the model at $\alpha = 0.05$. Moreover, R^2_{adj}=0.9673, means that 96.73% of the variation of the response variable is explained by the interation model compared to only 0.8962 for the additive model that predicts sales using TV and radio without an interaction term.

Note that from the additive model assume that the effect on sales of TV advertising is independent of the effect of radio advertising. This assumption might not be true. For example, spending money on TV advertising may increase the effectiveness of radio advertising on sale. In marketing, this is known as **a synergy effect**, in statistics it is referred to as **an interaction effect**.

Interpreting Coefficients of Predictor Variables

Notice that the model also can be rewritten as

$$\begin{aligned}
Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\
&= \beta_0 + \phi X_1 + \beta_2 X_2 + \epsilon \\
where & \\
\phi &= \beta_1 + \beta_3 X_2.
\end{aligned}$$

Since ϕ changes with X_2, the effect of X_1 on Y is no longer constant: adjusting X_2 will change the impact of X_1 on Y. The model also can be rewritten as

$$\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon \\
&= \beta_0 + \beta_1 X_1 + \phi X_2 + \epsilon \\
where & \\
\phi &= \beta_2 + \beta_3 X_1
\end{aligned}$$

For example, suppose that we are interested in studying the productivity of a factory. We wish to predict the number of units produced on the basis of the number of production lines and the total number of workers. It seems likely that the effect of increasing the number of production lines will depend on the number of workers, since if no workers are available to operate the lines, then increasing the number of lines will not increase production. This suggests that it would be appropriate to include an interaction term between lines and workers in a linear model to predict units. Suppose that when we fit the model, we obtain

$$\begin{aligned}\hat{units} &= 1.2 + 3.4lines + 0.22workers + 1.4(lines \cdot workers)\\ &= 1.2 + (3.4 + 1.4workers)lines + 0.22workers\end{aligned}$$

In other words, adding an additional line will increase the average number of units produced by 3.4 + 1.4*workers. Hence, the more workers we have, the stronger will be the effect of lines. Let's return to the Advertising example. A linear model that uses radio, TV, and an interaction between the two to predict sales takes the form

$$\begin{aligned}Sale &= \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3(TV \cdot radio) + \epsilon\\ &= \beta_0 + (\beta_1 + \beta_3 radio)TV + \beta_2 radio + \epsilon\end{aligned}$$

We can interpret the coefficient $\beta_1 + \beta_3 radio$ as: spending additional 1,000 dollars on TV advertising leads to an *increase* in sales by approximately $\beta_1 + \beta_3 radio$ units.

The results from the output strongly suggest that the model that includes the interaction term is superior to the model that contains only main effects

The coefficient estimates in the output suggest that an increase in TV advertising of 1,000 dollars is associated with increased sales of $(\beta_1 + \beta_3 radio) \times 1000 = 19 + 1.1radio$ units. And an increase in radio advertising of 1,000 dollars will be associated with an increase in sales of $(\beta_2 + \beta_3 TV) \times 1,000 = 29 + 1.1TV)$ units.

In this example, the p-values associated with TV, radio, and the interaction term all are statistically significant and so it is obvious that all three variables should be included in the model. However, it is sometimes the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not.

The hierarchical principle states that if we include an interaction in a model, we should also include the main effects,even if the p-values associated with principle their coefficients are not significant.

Caution That is, if the interaction between X_1 and X_2 seems important, then we should include both X_1 and X_2 in the model even if their coefficient estimates have large p-values.

The rationale for this principle is that if $X_1 \times X_2$ is related to the response,then whether or not the coefficients of X$_1$ or X$_2$ are exactly zero is of little interest. Also $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

Inclass Pratice Problem 4

From the condominium problem, do the data provide sufficient evidence to indicate that the interation term need to be added in the model? If you had to compare additive models with the interaction model, which model would you choose? Explain

In-class Practice Problem 5

Data on last year's sale (Y in 100,000s dollars) in 40 sales districts are given in the sales.csv file. This file also contains

promotional expenditures (X_1: in 1,000s dollars),

the number of active accounts (X_2),

the number of competing brands (X_3) and

the district potential (X_4, coded) for each of the district (OMIT THIS VARIABLE FOR NOW)

1. Find the best fit additive to predict sales using some or all of the variables X_1, X_2, X_3 only.
2. Find the best fit model with interaction terms (if needed) using some or all of the variables X_1, X_2, X_3
3. Which model would you choose? Explain.

4. Once you obtain the best fit model, interpret the regression coefficient for X_3 (Hint: it will interact with another variable).

Multiple Regression with Qualitative (Dummy) Variable Models

Multiple regression models can also be written to include qualitative (or categorical) independent variables. Qualitative variables, unlike quantitative variables, cannot be measured on a numerical scale. Therefore, we must code the values of the qualitative variable (called levels) as numbers before we can fit in the model. These coded qualitative variables are called **dummy variables** or **categorical predictor variables**, since the number assigned to the various levels are arbitrarily selected.

Dummy Coding

Because categorical predictor variables cannot be entered directly into a regression model and be meaningfully interpreted, some other methods of dealing with information of this type must be developed. In general, a categorical variable with k levels will be transformed into $k-1$ variables each with two levels. For example, if a categorical variable had six levels, then five dichotomous variables could be constructed that would contain the same information as the single categorical variable. The process of creating dichotomous variables from categorical variables is called **dummy coding**.

Dummy Coding with two levels The simplest case of dummy coding is when a categorical variable has two levels by assigning zero and one to the variable.

For example, the Credit data set records balance (average credit card debt for a number of individuals) as well as several quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating). In addition to these quantitative variables, we also have four qualitative variables: gender, student (student status), status (marital status), and ethnicity (Caucasian, African American or Asian). Data are provided in **credit.csv**_ file. Suppose that we wish to investigate differences in credit card balance between males and females. Based on the gender variable, we can create a dummy variable with 0 as male and 1 as female.

```
credit=read.csv("credit.csv",header = TRUE)
head(credit)
```

```
##   number  Income Limit Rating Cards Age Education Gender Student Married
## 1      1  14.891  3606    283     2  34        11   Male      No     Yes
## 2      2 106.025  6645    483     3  82        15 Female     Yes     Yes
## 3      3 104.593  7075    514     4  71        11   Male      No      No
## 4      4 148.924  9504    681     3  36        11 Female      No      No
## 5      5  55.882  4897    357     2  68        16   Male      No     Yes
## 6      6  80.180  8047    569     4  77        10   Male      No      No
##   Ethnicity Balance
## 1 Caucasian     333
## 2     Asian     903
## 3     Asian     580
## 4     Asian     964
## 5 Caucasian     331
## 6 Caucasian    1151
```

```
dummymodel<-lm(Balance~factor(Gender),data=credit)
summary(dummymodel)
```

```
##
## Call:
## lm(formula = Balance ~ factor(Gender), data = credit)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -529.54 -455.35  -60.17  334.71 1489.20
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            509.80      33.13  15.389   <2e-16 ***
## factor(Gender)Female    19.73      46.05   0.429    0.669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.2 on 398 degrees of freedom
## Multiple R-squared:  0.0004611,  Adjusted R-squared:  -0.00205
## F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

```r
#option2
dummymodel1<-lm(Balance~Gender,data=credit)
summary(dummymodel1)
```

```
##
## Call:
## lm(formula = Balance ~ Gender, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -529.54 -455.35  -60.17  334.71 1489.20
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    509.80      33.13  15.389   <2e-16 ***
## GenderFemale    19.73      46.05   0.429    0.669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.2 on 398 degrees of freedom
## Multiple R-squared:  0.0004611,  Adjusted R-squared:  -0.00205
## F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

Rfunction

factor() : command will make sure that R knows that your variable is categorical. This is especially useful if your categories are indicated by integers, otherwise function lm() might interpret the variable as continuous.

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \epsilon$$

$$\widehat{balance}_i = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if } i^{th} \text{ person is female} \\ \beta_0 + \epsilon & \text{if } i^{th} \text{ person is male} \end{cases}$$

Interpreting Coefficients of Predictor Variables

β_0 can be interpreted as the average credit card balance among males,

β_1 as the average difference in credit card balance between females and males.

$\beta_0 + \beta_1$ can be interpreted as the average credit card balance among females.

$$\hat{Y_i} = 509.80 + 19.73X_{1i}$$

$$\widehat{balance_i} = \begin{cases} 509.80 + 19.73 = 529.53 & \text{if } i^{th} \text{ person is female} \\ 509.80 & \text{if } i^{th} \text{ person is male} \end{cases}$$

From the output, the coeffcient estimates and other information associated with the model are provided. The average credit card debt for males is estimated to be 509.80 dollars whereas females are estimated to carry 19.73 in additional debt for a total of 509.80 + 19.73 = 529.53 dollars. However, we notice that the p-value for the dummy variable is very high. This indicates that there is no statistical evidence of a difference in average credit card balance between the genders.

Inclass Practice Problem 6

Suppose that we wish to investigate differences in credit card balance between marital status. Based on the Married variable, we can create a dummy variable which 0 is NO and 1 is Yes. Create a simple linear regression model to predict the credit card balance by using the Married variable. What are the regression coefficients for this model? How do you interpret the regression coefficients? **Ignore the individual t-test output**

Dummy Coding with three levels When the categorical variable has three levels, it is converted to two dichotomous (dummy) variables.

For example, there is always a certain curiosity and controversy surrounding professor' salaries and whether they are overpaid or not paid enough. A university would like to study the effects of ranks and department on the salaries. 30 observations were randomly chosen from 3 different departments. The data are provided in **salary.csv** data file.

gender= (0=Male, 1=Female)

rank= (1=Assistant, 2=Associate, 3=Full)

Dept= Department (1=Family Studies, 2=Biology, 3=Business)

year=Years since making Rank

merit=Average Merit Ranking

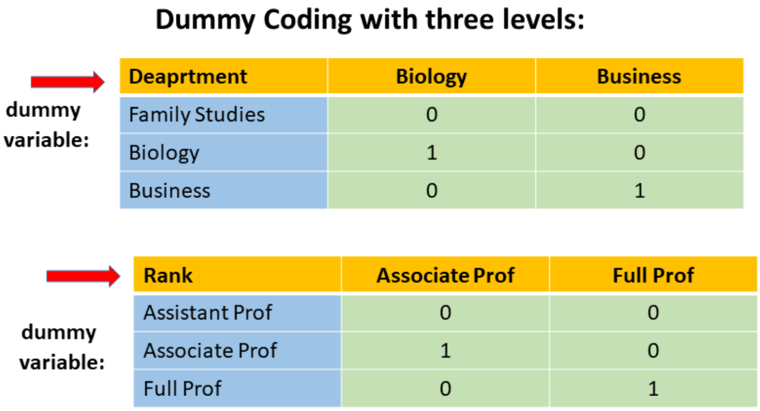
The variable *Dept* has three levels: 1=Family Studies, 2=Biology, and 3=Business. Variable *Dept* could be dummy coded into two variables, one called Biology and one called Business. The variable *rank* has also three levels and will be also coded into two dummy variables: Assistant Prof and Full Prof. The dummy coding is represented below.

Before considering both predictors, let practice how to interpret the regression coefficients for each categorical variable.

For example, considering only rank variable with three levels

```
salary=read.csv("salary.csv",header = TRUE)
head(salary)
```

```
##   salary gender rank dept year merit
## 1     38      0    3    1    0  1.47
## 2     58      1    2    2    8  4.38
## 3     80      1    3    2    9  3.65
## 4     30      1    1    1    0  1.64
## 5     50      1    1    3    0  2.54
## 6     49      1    1    3    1  2.06
```



Dummy Coding with three levels:

dummy variable:

Deaprtment	Biology	Business
Family Studies	0	0
Biology	1	0
Business	0	1

dummy variable:

Rank	Associate Prof	Full Prof
Assistant Prof	0	0
Associate Prof	1	0
Full Prof	0	1

Figure 1: Dummy Coding with three levels

```
dummymodel<-lm(salary~factor(rank),data=salary)
summary(dummymodel)
```

```
##
## Call:
## lm(formula = salary ~ factor(rank), data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.875  -5.799   0.000   5.353  23.125
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     42.000      2.259  18.593  < 2e-16 ***
## factor(rank)2   10.571      4.005   2.640 0.013613 *
## factor(rank)3   14.875      3.830   3.884 0.000602 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.749 on 27 degrees of freedom
## Multiple R-squared:  0.3881, Adjusted R-squared:  0.3428
## F-statistic: 8.563 on 2 and 27 DF,  p-value: 0.001319
```

```
#forget factor(...)
dummymodel1<-lm(salary~rank,data=salary)
summary(dummymodel1)
```

```
##
## Call:
```

```
## lm(formula = salary ~ rank, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.9017  -5.2168   0.1139   5.8639  22.0983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.871      3.684   9.466 3.19e-10 ***
## rank           7.677      1.882   4.080 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.697 on 28 degrees of freedom
## Multiple R-squared:  0.3729, Adjusted R-squared:  0.3505
## F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003389
```

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon$$

$$salary_i = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if } i^{th} \text{ person is ranked as Associate Prof} \\ \beta_0 + \beta_2 + \epsilon & \text{if } i^{th} \text{ person is ranked as Full Prof} \\ \beta_0 + \epsilon & \text{if } i^{th} \text{ person is ranked as Assistant Prof} \end{cases}$$

```
salary=read.csv("salary.csv",header = TRUE)
dummymodel<-lm(salary~factor(rank),data=salary)
summary(dummymodel)
```

```
##
## Call:
## lm(formula = salary ~ factor(rank), data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.875  -5.799   0.000   5.353  23.125
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     42.000      2.259  18.593  < 2e-16 ***
## factor(rank)2   10.571      4.005   2.640 0.013613 *
## factor(rank)3   14.875      3.830   3.884 0.000602 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.749 on 27 degrees of freedom
## Multiple R-squared:  0.3881, Adjusted R-squared:  0.3428
## F-statistic: 8.563 on 2 and 27 DF,  p-value: 0.001319
```

Interpreting Coefficients of Predictor Variables

β_0 can be interpreted as the average salary for Assistant Professor position ,

β_1 as the difference in average salary between Associate Professor and Assistant Professor.

β_2 as the difference in average salary between Full Professor and Assistant Professor.

$\beta_0 + \beta_1$ can be interpreted as the average salary for Associate Professor position .

$\beta_0 + \beta_2$ can be interpreted as the average salary for Full Professor position .

In-class Practice Problem 7

There is always a certain curiosity and controversy surrounding professors' salaries and whether they are overpaid or not paid enough. A university would like to study the effects of ranks and departments on salaries. 30 observations were randomly chosen from 3 different departments. The data are provided in the salary.csv data file. Dept= Department (1=Family Studies, 2=Biology, 3=Business)

Instead of the rank variable, practice how to interpret the dept variable.

Example: There is always a certain curiosity and controversy surrounding professor'salaries and whether they are overpaid or not paid enough. A university would like to study the effects of ranks and dept on the salaries (thousands of dollars). 30 observations were randomly chosen from 3 different departments. The data are provided in **salary.csv** data file.

```r
salary=read.csv("salary.csv",header = TRUE)
dummymodel<-lm(salary~factor(rank)+factor(dept),data=salary)
summary(dummymodel)
```

```
##
## Call:
## lm(formula = salary ~ factor(rank) + factor(dept), data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.243  -3.333  -0.049   2.350  20.256
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     34.049      2.308  14.754 7.62e-14 ***
## factor(rank)2   13.208      2.983   4.427 0.000164 ***
## factor(rank)3   15.194      2.797   5.433 1.22e-05 ***
## factor(dept)2   10.502      2.972   3.533 0.001624 **
## factor(dept)3   13.351      2.752   4.851 5.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.368 on 25 degrees of freedom
## Multiple R-squared:  0.6998, Adjusted R-squared:  0.6518
## F-statistic: 14.57 on 4 and 25 DF,  p-value: 2.856e-06
```

Rfunction

factor() : command will make sure that R knows that your variable is categorical. This is especially useful if your categories are indicated by integers, otherwise function lm() will interpret the variable as continuous.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon$$

$$salary_i = \begin{cases} \beta_0+\beta_1+\epsilon & \text{if } i^{th}\text{person is ranked as Associate Prof and is from Family Studies dept} \\ \beta_0+\beta_2+\epsilon & \text{if } i^{th}\text{person is ranked as Full Prof and is from Family Studies dept} \\ \beta_0+\beta_3+\epsilon & \text{if } i^{th}\text{person is ranked as Assistant Prof and is from Biology Dept} \\ \beta_0+\beta_4+\epsilon & \text{if } i^{th}\text{person is ranked as Assistant Prof and is from Business Dept} \\ \beta_0+\beta_1+\beta_3+\epsilon & \text{if } i^{th}\text{person is ranked as Associate Prof and is from Biology dept} \\ \beta_0+\beta_1+\beta_4+\epsilon & \text{if } i^{th}\text{person is ranked as Associate Prof and is from Business dept} \\ \beta_0+\beta_2+\beta_3+\epsilon & \text{if } i^{th}\text{person is ranked as Full Prof and is from Biology dept} \\ \beta_0+\beta_2+\beta_4+\epsilon & \text{if } i^{th}\text{person is ranked as Full Prof and is from Business dept} \\ \beta_0+\epsilon & \text{if } i^{th}\text{person is ranked as Assistant Prof and is from Family Studies dept} \end{cases}$$

Interpreting Coefficients of Predictor Variable

β_0 can be interpreted as the average salary of an Assistant Prof from Family Studies dept.

β_1 can be interpreted as the average difference in salary between an Assistant Prof and an Associate Prof (hold dept)

β_2 can be interpreted as the average difference in salary between an Assistant Prof and a Full Prof (hold dept) .

.

.

Interaction Effect in Multiple Regression with both Quantitative and Qualitative (Dummy) Variable Models

In previous topics, we considered Multiple Regression models for both quantitative and qualitative variables. We also discussed an interaction in Multiple Regression for quantitative variables. However, the concept of interactions applies just as well to qualitative variables, or to a combination of quantitative and qualitative variables. In fact, an interaction between a qualitative variable and a quantitative variable has a particularly nice interpretation.

Consider the Credit data set example and suppose that we wish to predict balance using the income (quantitative) and student (qualitative) variables. In the absence of an interaction term, the model takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon$$

$$balance_i = \beta_0 + \beta_1 Income_i + \begin{cases} \beta_2 & \text{if } i^{th}\text{person is a student} \\ 0 & \text{if } i^{th}\text{person is not a student} \end{cases}$$

$$balance_i = \beta_1 Income_i + \begin{cases} \beta_0+\beta_2 & \text{if } i^{th}\text{person is a student} \\ \beta_0 & \text{if } i^{th}\text{person is not a student} \end{cases}$$

```r
credit=read.csv("credit.csv",header = TRUE)
mixmodel<-lm(Balance~Income+factor(Student), data=credit)
summary(mixmodel)
```

```
##
## Call:
## lm(formula = Balance ~ Income + factor(Student), data = credit)
##
```

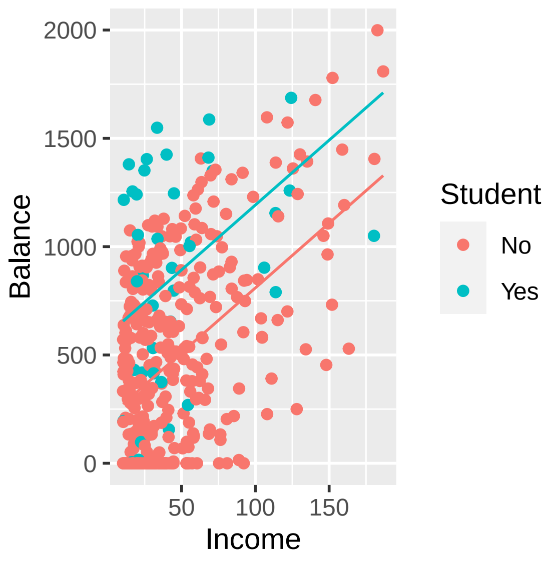
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -762.37 -331.38  -45.04  323.60  818.28
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         211.1430    32.4572   6.505 2.34e-10 ***
## Income                5.9843     0.5566  10.751  < 2e-16 ***
## factor(Student)Yes 382.6705    65.3108   5.859 9.78e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.8 on 397 degrees of freedom
## Multiple R-squared:  0.2775, Adjusted R-squared:  0.2738
## F-statistic: 76.22 on 2 and 397 DF,  p-value: < 2.2e-16
```

$$balance_i = 5.9843 Income_i + \begin{cases} 211.1430 + 382.6705 = 593.8135 & \text{if } i^{th}\text{person is a student} \\ 211.1430 & \text{if } i^{th}\text{person is not a student} \\ \end{cases}$$

$$balance_i = \begin{cases} 593.8135 + 5.9843 Income_i & \text{if } i^{th} \text{ person is a student} \\ 211.1430 + 5.9843 Income_i & \text{if } i^{th} \text{ person is not a student} \end{cases}$$

Notice that this amounts to fitting two parallel lines to the data, one for students and one for non-students. The lines for students and non-students have different intercepts, $\beta_0 + \beta_2$ versus β_0, but the same slope, β_1. This is illustrated in the plot below. The fact that the lines are parallel means that the average effect on balance of a one-unit increase in income does not depend on whether or not the individual is a student. **This represents a potentially serious limitation of the model, since in fact a change in income may have a very different effect on the credit card balance**

```
library(ggplot2)
credit=read.csv("credit.csv",header = TRUE)
mixmodel<- lm(Balance~Income+factor(Student),data=credit)
#For student y=593.8135+5.9843Income
#For nonstudent  y=211.1430+5.9843 Income
nonstudent=function(x){coef(mixmodel)[2]*x+coef(mixmodel)[1]}
student=function(x){coef(mixmodel)[2]*x+coef(mixmodel)[1]+coef(mixmodel)[3]}
ggplot(data=credit,mapping= aes(x=Income,y=Balance,colour=Student))+geom_point()+
  stat_function(fun=nonstudent,geom="line",color=scales::hue_pal()(2)[1])+
  stat_function(fun=student,geom="line",color=scales::hue_pal()(2)[2])
```



This limitation can be addressed by adding an interaction variable, created by multiplying income with the dummy variable for student. Our model now becomes

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon$$

$$balance_i = \beta_0 + \beta_1 x Income_i + \begin{cases} \beta_2 + \beta_3 x Income_i & \text{if } i^{th} \text{ person is a student} \\ 0 & \text{if } i^{th} \text{ person is not a student} \end{cases}$$

$$balance_i = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x Income_i & \text{if } i^{th} \text{ person is a student} \\ \beta_0 + \beta_1 x Income_i & \text{if } i^{th} \text{ person is not a student} \end{cases}$$

```r
credit=read.csv("credit.csv",header = TRUE)
mixmodel<- lm(Balance~Income+factor(Student)+Income*factor(Student),data=credit)
summary(mixmodel)
```

```
##
## Call:
## lm(formula = Balance ~ Income + factor(Student) + Income * factor(Student),
##     data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.39 -325.70  -41.13  321.65  814.04
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 200.6232    33.6984   5.953 5.79e-09 ***
## Income                        6.2182     0.5921  10.502  < 2e-16 ***
## factor(Student)Yes          476.6758   104.3512   4.568 6.59e-06 ***
## Income:factor(Student)Yes  -1.9992     1.7313  -1.155    0.249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.6 on 396 degrees of freedom
## Multiple R-squared:  0.2799, Adjusted R-squared:  0.2744
## F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

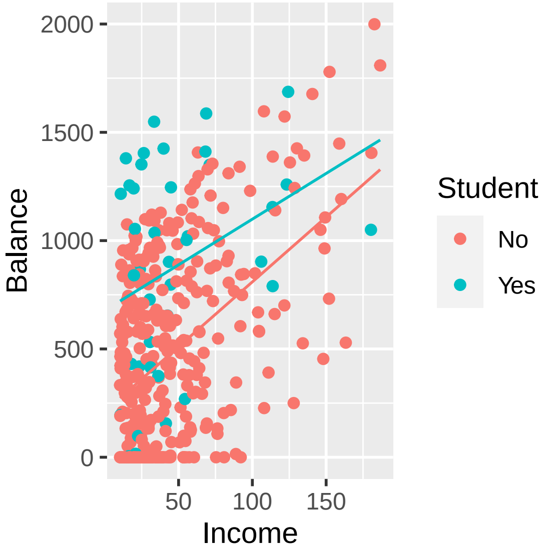
$$Y_i = 200.6232 + 6.2182X_{i1} + 476.6758X_{i2} - 1.9992X_{i1}X_{i2} + \epsilon$$

$$balance_i = 200.6232 + 6.2182Income_i + \begin{cases} 476.6758 - 1.9992Income_i & \text{if } i^{th}\text{person is a student} \\ 0 & \text{if } i^{th}\text{person is not a student} \end{cases}$$

$$\widehat{balance_i} = \begin{cases} (200.6232 + 476.67582) + (6.2182 - 1.9992)Income_i & \text{if } i^{th} \text{ person is a student} \\ 200.6232 + 6.2182Income_i & \text{if } i^{th} \text{ person is not a student} \end{cases}$$

$$\widehat{balance_i} = \begin{cases} 677.29902 + 4.219Income_i & \text{if } i^{th} \text{ person is a student} \\ 200.6232 + 6.2182Income_i & \text{if } i^{th} \text{ person is not a student} \end{cases}$$

```
library(ggplot2)
credit=read.csv("credit.csv",header = TRUE)
mixmodel<- lm(Balance~Income+factor(Student),data=credit)
#For student y=677.2992+4.219Income
#For nonstudent  y=200.6232+6.2182Income
student=function(x){4.219*x+677.2992}
nonstudent=function(x){coef(mixmodel)[2]*x+coef(mixmodel)[1]}
ggplot(data=credit,mapping= aes(x=Income,y=Balance,colour=Student))+geom_point()+
  stat_function(fun=nonstudent,geom="line",color=scales::hue_pal()(2)[1])+
  stat_function(fun=student,geom="line",color=scales::hue_pal()(2)[2])
```



Disregard the p-value for the interaction term, we have two different regression lines for the students and the non-students. But now those regression lines have different intercepts, $\beta_0 + \beta_2$ versus β_1, as well as different slopes, $\beta_1 + \beta_3$ versus β_1. This allows for the possibility that changes in income may affect the credit card balances of students and non-students differently. The output shows the estimated relationships between income and balance for students and non-students in the model. We note that the slope for students (4.219) is lower than the slope for non-students (6.218). This suggests that increases in income are associated with smaller increases in credit card balance among students as compared to non-students.

Inclass Practice Problem 8

From the credit card example, use the lm() function to perform the best fit model. How would you interpret the regression coeffients (if possible)? Would you recommend this model for predictive purpose?

From the inclass practice problem, you can see that it is quite complicated (possible) to interpret regression coefficients as there are so many predictors in the model. However, let's practice the example below.

```
credit=read.csv("credit.csv",header = TRUE)
mixmodel2<-lm(Balance~Rating+Income+factor(Student)+
                factor(Student)*Rating+factor(Student)*Income+
                Rating*Income,data=credit)
summary(mixmodel2)
```

```
##
## Call:
## lm(formula = Balance ~ Rating + Income + factor(Student) + factor(Student) *
##     Rating + factor(Student) * Income + Rating * Income, data = credit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -207.799  -74.552   -3.999   74.055  253.340
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -4.712e+02  2.053e+01 -22.945  < 2e-16 ***
## Rating                     3.701e+00  6.312e-02  58.642  < 2e-16 ***
## Income                    -9.925e+00  4.696e-01 -21.136  < 2e-16 ***
## factor(Student)Yes         1.862e+02  4.442e+01   4.193 3.41e-05 ***
## Rating:factor(Student)Yes  1.031e+00  1.717e-01   6.003 4.40e-09 ***
## Income:factor(Student)Yes -2.775e+00  6.733e-01  -4.121 4.59e-05 ***
## Rating:Income              4.140e-03  7.098e-04   5.832 1.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.05 on 393 degrees of freedom
## Multiple R-squared:  0.957,  Adjusted R-squared:  0.9564
## F-statistic:  1458 on 6 and 393 DF,  p-value: < 2.2e-16
```

$$\begin{aligned} balance &= \beta_0 + \beta_1 Rating + \beta_2 Income + \beta_3 Student \\ &\quad + \beta_4 Rating * Student + \beta_5 Income * Student + \beta_6 Rating * Income + \epsilon \end{aligned}$$

if a person is a student

$$\begin{aligned} &balance_i = \beta_0 + \beta_1 Rating + \beta_2 Income + \beta_3(1) + \beta_4 Rating * (1) \\ &+ \beta_5 Income * (1) + \beta_6 Rating * Income + \epsilon \end{aligned}$$

if a person is not a student

$$
\begin{aligned}
&balance_i = \beta_0 + \beta_1 Rating + \beta_2 Income \\
&+ \beta_6 Rating * Income + \epsilon
\end{aligned}
$$

A Quadratic (Second-Order) Model with Quantitative Predictors

All of the models discussed in the previous sections proposed straight-line relationships between $E(y)$ and each of the independent variables in the model. In this section, we consider a model that allows for curvature in the relationship. This model is a second-order model because it will include an X^2 term. Here, we consider a model that includes only one independent variable X. The form of this model, called the *quadratic model*, is

$$
\begin{aligned}
Y =& \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \epsilon \\
\hat{y} =& \hat{\beta_0} + \hat{\beta_1} X_1 + \hat{\beta_2} X_1^2
\end{aligned}
$$

The term involving X_1^2, called a quadratic term (or second-order term), enables us to hypothesize curvature in the graph of the response model relating Y to X_1. Graphs of the quadratic model for two different values of β_2 are shown in the figure below. When the curve opens upward, the sign of β_2 is positive (see Figure 2 (a); when the curve opens downward, the sign of β_2 is negative (see Figure 2 (b)).

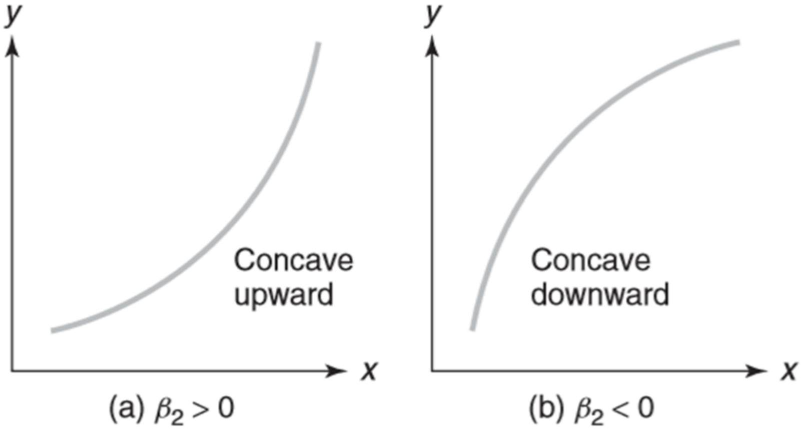


Figure 2: Graphs for two quadratic models

Interpretation of the regression coefficients

The interpretation of the estimated coefficients in a quadratic model must be under taken cautiously.

$\hat{\beta_0}$ can be meaningfully interpreted only if the range of the independent variable includes zero-that is, if $X_1 = 0$ is included in the sampled range of X_1.

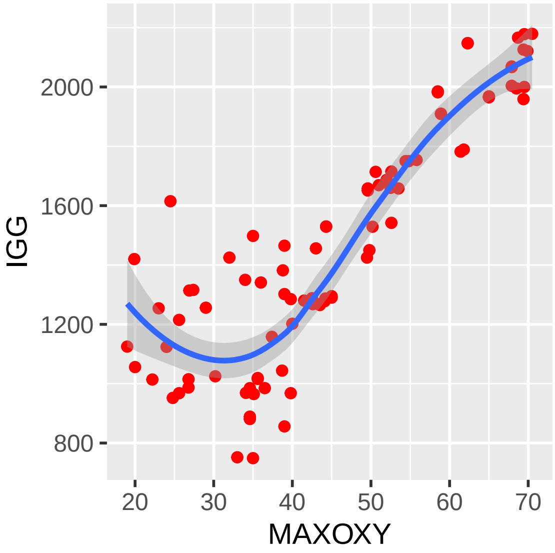
$\hat{\beta_1}$ no longer represents a slope in the presence of the quadratic term X_1^2. The estimated coefficient of the first-order term X_1 will not, in general, have a meaningful interpretation in the quadratic model.

$\hat{\beta}_2$, the sign of the coefficients, $\hat{\beta}_2$, of the quadratic term, X_1^2, is the indicator of whether the curve is concave downward (mound-shaped) or concave upward (bowl-shaped). A negative $\hat{\beta}_2$ implies downward concavity, as in this example (Figure 2), and a positive $\hat{\beta}_2$ implies upward concavity. Rather than interpreting the numerical value of $\hat{\beta}_2$ itself.

Example A physiologist wants to investigate the impact of exercise on the human immune system. The physiologist theorizes that the amount of immunoglobulin Y in blood (called IgG, an indicator of long-term immunity, milligrams) is related to the maximal oxygen uptake x (a measure of aerobic fittness level, milliliters per kilogram). The data file is provided in **AEROBIC.CSV** file. Construct a scatterplot for the data. Is there evidence to support the use of a quadratic model? What is the best model to fit the data.

```
library(ggplot2)  #using ggplot2 for data visualization
aerobicdata=read.csv("AEROBIC.csv",header = TRUE)
ggplot(data=aerobicdata,mapping= aes(x=MAXOXY,y=IGG))+geom_point(color='red')+
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
simplemodel=lm(IGG~MAXOXY,data=aerobicdata)
summary(simplemodel)
```

```
##
## Call:
## lm(formula = IGG ~ MAXOXY, data = aerobicdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -478.11 -127.30   28.04  116.38  636.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  398.954     69.561   5.735 1.38e-07 ***
## MAXOXY        23.662      1.468  16.120  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 201.4 on 87 degrees of freedom
## Multiple R-squared:  0.7492, Adjusted R-squared:  0.7463
## F-statistic: 259.8 on 1 and 87 DF,  p-value: < 2.2e-16
```

```
quadmodel=lm(IGG~MAXOXY+I(MAXOXY^2),data=aerobicdata)
summary(quadmodel)
```

```
##
## Call:
## lm(formula = IGG ~ MAXOXY + I(MAXOXY^2), data = aerobicdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -439.91  -86.43  -30.15  139.15  517.61
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1270.41137  186.19900   6.823 1.18e-09 ***
## MAXOXY       -18.10744    8.52049  -2.125   0.0364 *
## I(MAXOXY^2)    0.45082    0.09088   4.960 3.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 178.6 on 86 degrees of freedom
## Multiple R-squared:  0.805,  Adjusted R-squared:  0.8004
## F-statistic: 177.5 on 2 and 86 DF,  p-value: < 2.2e-16
```

```
cubemodel=lm(IGG~MAXOXY+I(MAXOXY^2)+I(MAXOXY^3),data=aerobicdata)
summary(cubemodel)
```

```
##
## Call:
## lm(formula = IGG ~ MAXOXY + I(MAXOXY^2) + I(MAXOXY^3), data = aerobicdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -356.7 -100.1  -12.5  103.6  496.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.502e+03  5.015e+02   6.982 6.03e-10 ***
## MAXOXY      -1.902e+02  3.727e+01  -5.103 2.01e-06 ***
## I(MAXOXY^2)  4.527e+00  8.680e-01   5.216 1.27e-06 ***
## I(MAXOXY^3) -2.999e-02  6.357e-03  -4.717 9.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 159.9 on 85 degrees of freedom
## Multiple R-squared:  0.8454, Adjusted R-squared:   0.84
## F-statistic:   155 on 3 and 85 DF,  p-value: < 2.2e-16
```

```
forthmodel=lm(IGG~MAXOXY+I(MAXOXY^2)+I(MAXOXY^3)+I(MAXOXY^4),data=aerobicdata)
summary(forthmodel)# should stop at cubemodel because all variables are not significant.
```

```
##
## Call:
## lm(formula = IGG ~ MAXOXY + I(MAXOXY^2) + I(MAXOXY^3) + I(MAXOXY^4),
##     data = aerobicdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -362.89 -104.07   -8.92   98.60  481.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.012e+03  1.596e+03   1.261    0.211
## MAXOXY      -3.370e+01  1.635e+02  -0.206    0.837
## I(MAXOXY^2) -1.255e+00  5.947e+00  -0.211    0.833
## I(MAXOXY^3)  5.979e-02  9.156e-02   0.653    0.516
## I(MAXOXY^4) -4.976e-04  5.063e-04  -0.983    0.328
##
## Residual standard error: 160 on 84 degrees of freedom
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8399
## F-statistic: 116.4 on 4 and 84 DF,  p-value: < 2.2e-16
```

R function

I(X^2) :add quadratic term to the model

From the output, considering the scatterplot between y and X, we found that the best model to fit the data is

$$\hat{Y} = 1270.41137 - 18.10744X_1 + 0.45082X_1^2$$

moreover, $R^2_{adj} = 0.805$ and RMSE=178.6,comparing to the simple linear model, we can conclude that the quadratic model fits the data better than the simple linear regression model.

Note!

Model interpretations are not meaningful outside the range of the independent variable. Although the model appears to support the data. To make a prediction for Y, value of X should be inside the range of the independent variable. Otherwise the prediction will not be meaningful.

Inclass Practice Problem 9

Suppose you wanted to model the quality, y, of a product as a function of the pressure pounds per square inch (psi), at which it is produced. Four inspectors independently assign a quality score between 0 and 100 to each product, and then the quality, y, is calculated by averaging the four scores. An experiment is conducted by varying temperature in F. Fit a second-order model to the data and sketch the scatterplot. The data are provided in **PRODQUAL.csv** file

Exercise 2

The amount of water used by the production facilities of a plant varies. Observations on water usage and other,possibility related,variables were collected for 250 months. The data are given in **water.csv file** The explanatory variables are

TEMP= average monthly temperature(degree celsius)

PROD=amount of production(in hundreds of cubic)

DAYS=number of operationing day in the month (days)

HOUR=number of hours shut down for maintenance (hours)

The response variable is USAGE=monthly water usage (gallons/minute)

From the best model that you analysed in Exercise 1, build an interaction model to fit the multiple regression model. From the output, which model would you recommend for predictive purposes? Interpret the regression coefficients.

References

-*Gareth James & Daniela Witten & Trevor Hastie Robert Tibshirani, An Introduction to Statistical Learning with Applications in R: Springer New York Heidelberg Dordrecht London.*

-*Wickham and Grolemund, R for Data Science: O'Reilly Media*