



筑波大学

*University of Tsukuba*

# Multiple Linear Regression Model

Marco Antonio Florenzano Mollinetti<sup>1</sup>

<sup>1</sup>**University of Tsukuba, Systems Optimization Laboratory**

[mollinetti@syoun.cs.tsukuba.ac.jp](mailto:mollinetti@syoun.cs.tsukuba.ac.jp)




# Before we Begin

- Go to the github repo:
  - <https://github.com/Mollinetti/Experiment-Design-R>
  - Download the script for this class! (in the 'scripts' folder, class\_6.r!)
- Run the snippet at the beginning to load/install the required libraries



# Agenda

- Introduction
- Multiple Linear Regression
- Interaction Terms
- Dummy variables
- Verifying collinearity
- Using many predictors for the fit

- 
- Introduction
  - Multiple Linear Regression
  - Interaction Terms
  - Dummy variables
  - Verifying collinearity
  - Using many predictors for the fit



# Introduction

- Multiple linear regression uses more than one predictor in  $X$
- We cannot plot the predictor against response anymore
- We now have to check for collinearity in our assumptions




# Introduction

- We still follow the **t-statistic** and **F-test** results
- **R-squared** and **RSS** are not entirely reliable anymore
- **AIC** and **BIC** are more recommended  
(however they do not give an estimate of how good the fit is)



# Introduction

- Load the “Healthy\_breakfast.csv” dataset
- Columns:

- 
- Introduction
  - Multiple Linear Regression
  - Interaction Terms
  - Dummy variables
  - Verifying collinearity
  - Using many predictors for the fit



# Multiple Linear Regression

- Consider the data to have  $n$  data points and  $p$  predictors
- Multiple linear regression is recommended when  $n > p$  and  $n > 30$
- **NEVER use standard linear regression when  $p > n$**

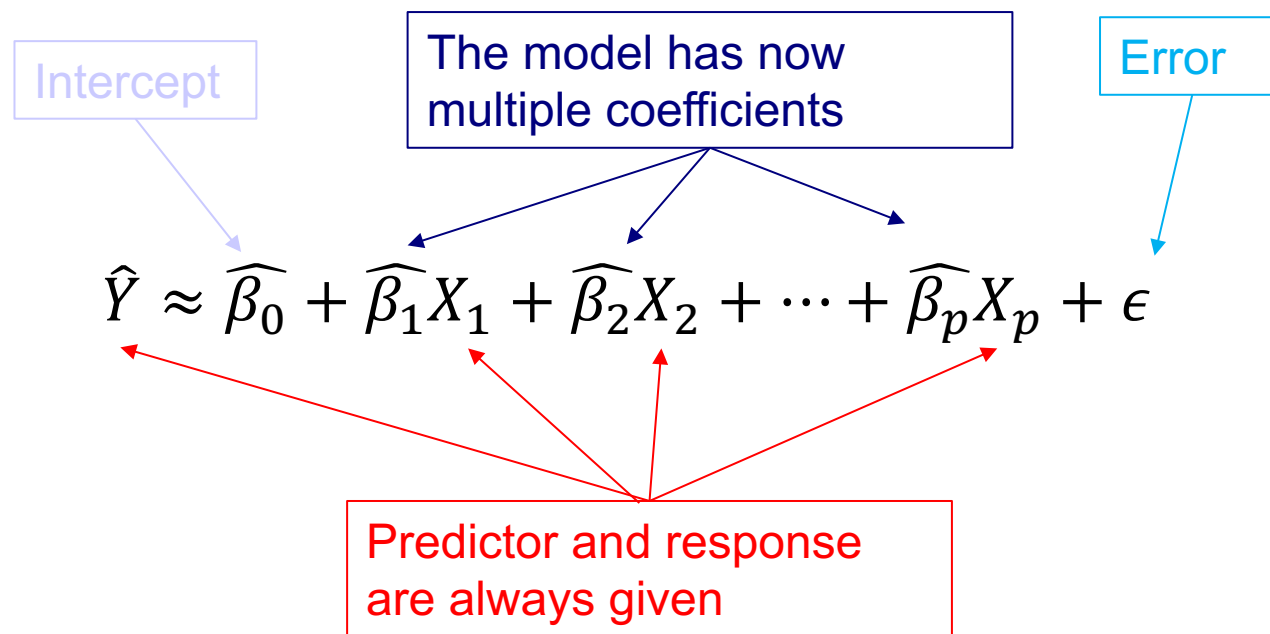
# Multiple Linear Regression

- The linear model has the following structure:

$$\hat{Y} \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_n X_n + \epsilon$$

# Multiple Linear Regression

- The linear model has the following structure:



# Multiple Linear Regression

- The coefficients are calculated the same way, using a least squares approach to minimize the residual squared error

$$\begin{aligned}\text{minimize}_{y \in \mathbb{R}} \text{RSS}(y) &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots \hat{\beta}_p x_{ip})^2\end{aligned}$$



# Multiple Linear Regression

- In the MLR setting, each coefficient represents a average effect of increase **while fixing the others**
- The **p-value of the t-statistic** points whether the coefficient has any relationship to the predictor



# Multiple Linear Regression

- When performing MLR, we are interested in answering some questions
  1. Is at least one of the predictors useful in predicting the response?
  2. Do all predictors help to explain  $Y$  or only a subset?
  3. How well does the model fit the data?

# Multiple Linear Regression

Is at least one of the predictors useful in predicting the response

- To answer the first question, we test against the following hypothesis:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1: \text{at least one } \beta_j \text{ is nonzero}$$

- This test is done calculating the F-statistic
  - To reject  $H_0$  we expect values greater than 1
  - More reliable than the individual associated p values



# Multiple Linear Regression

Do all predictors help to explain  $Y$  or only a subset?

- Ideally, we would try many different models
- However, the number of possible models is  $2^p$
- For that we do variable selection (next episode)
  - Forward selection
  - Backward selection
  - Mixed selection



# Multiple Linear Regression

- How well does the model fit the data?
- We use RSE and  $R^2$  for simple regression, but they are not reliable for MLR
  - RSE and  $R^2$  scale with the number of predictors
- BIC and AIC are favored
- Graphical information is also welcomed

# Multiple Linear Regression

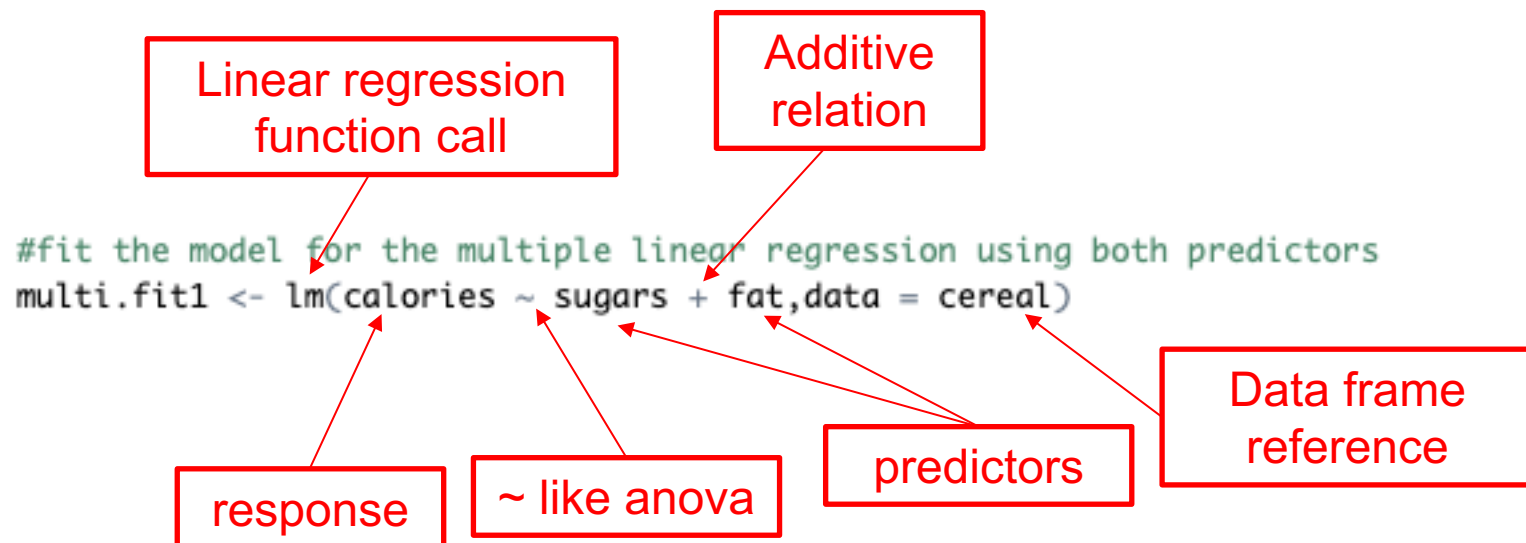
- MLR in R:

```
#fit the model for the multiple linear regression using both predictors  
multi.fit1 <- lm(calories ~ sugars + fat,data = cereal)
```



# Multiple Linear Regression


## ■ MLR in R:





# Multiple Linear Regression

- After building the MLR model, we need to:
  - Verify the t-statistic and F-test
  - Plot the residuals
  - Verify the assumptions
  - Estimate the goodness of fit
- A poor fit or fail of any assumption means that:
  - Other model is recommended
  - Data has to be transformed

- 
- Introduction
  - Multiple Linear Regression
  - Interaction Terms
  - Dummy variables
  - Verifying collinearity
  - Using many predictors for the fit



# Interaction Terms

- Two of the most important assumptions of MLR is that the relationship between  $X$  and  $Y$  must be additive and linear
- However, we can relax the additive assumption by adding an interaction term



# Interaction Terms

- Individual predictors may contribute to the explanation of the response
- The additive assumption may underestimate the predictor
- Inclusion of an extra predictor constructed by computing the product of  $X_i$  to  $X_j$

# Interaction Terms

- The linear model becomes the following:

$$\hat{Y} \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \dots + \beta_p X_p + \epsilon$$



# Interaction Terms

- The linear model becomes the following:

$$\hat{Y} \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \dots + \beta_p X_p + \epsilon$$

Interaction  
term



Interaction  
coefficient





# Interaction Terms

- The effect of each predictor is not constant anymore
- In this example, adjusting  $X_2$  will change the impact of  $X_1$  on  $Y$
- If the interaction between  $X_1$  and  $X_2$  seems important, we should include  $X_1$  and  $X_2$  in the model even if their coefficient estimates have large p-values!

# Interaction Terms

- In R, Adding an interaction term is straightforward

```
#fit the model for the multiple linear regression using both predictors and an interaction term
multi.fit2 <- lm(calories ~ sugars + fat + sugars:fat,data = cereal)

#the same thing can be achieved writing
multi.fit2 <- lm(calories ~ sugars * fat,data = cereal)
```



# Interaction Terms

- In R, Adding an interaction term is straightforward


Interaction  
between predictors

```
#fit the model for the multiple linear regression using both predictors and an interaction term  
multi.fit2 <- lm(calories ~ sugars + fat + sugars:fat,data = cereal)
```

```
#the same thing can be achieved writing  
multi.fit2 <- lm(calories ~ sugars * fat,data = cereal)
```

Same effect



- 
- Introduction
  - Multiple Linear Regression
  - Interaction Terms
  - Dummy variables
  - Verifying collinearity
  - Using many predictors for the fit



# Dummy Variables

- The concept of interaction applies to qualitative variables as well
- A dummy variable is a qualitative variable that has been converted to a quantitative variable
- A dummy variable can cover 2 levels, more are needed if more levels are needed

# Dummy Variables

- In our cereal example, the linear model with a quantitative (sugars) and qualitative (vitamin) would be:
  - Levels of Vitamin: {Enriched, 100%, None}

$$x_{i1} = \begin{cases} 1 & \text{if enriched} \\ 0 & \text{otherwise} \end{cases} \quad x_{i2} = \begin{cases} 1 & \text{if 100\%} \\ 0 & \text{otherwise} \end{cases}$$

$$y_i \approx \beta_0 + \beta_1 \text{sugars} + \beta_2 x_{i1} + \beta_3 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \beta_2 + \epsilon & \text{if enriched} \\ \beta_0 + \beta_1 + \beta_3 + \epsilon & \text{if 100\%} \\ \beta_0 + \beta_1 + \epsilon & \text{if none} \end{cases}$$



# Dummy Variables

- There will always be one less dummy variable than levels
- **None** is the level with no dummy variable, called **baseline**



# Dummy Variables

- In R, nothing else is needed to add a dummy variable to a linear model:

```
multi.fit3 <- lm(calories ~ sugars + vitamins ,data = cereal)
```




# Dummy Variables

- In R, nothing else is needed to add a dummy variable to a linear model:

```
multi.fit3 <- lm(calories ~ sugars + vitamins ,data = cereal)
```

Qualitative  
predictor



# Dummy Variables


- We can check the contrasts of the dummy variables

```
#contrasts of the vitamins  
contrasts(vitamins)
```

```
> contrasts(vitamins)  
      enriched none  
100%      0      0  
enriched  1      0  
none      0      1
```

In this case, 100%  
is the baseline!



- 
- Introduction
  - Multiple Linear Regression
  - Interaction Terms
  - Dummy variables
  - Verifying collinearity
  - Using many predictors for the fit



# Verifying Collinearity

- Two or more predictor variables are related to one another
- Can be difficult to separate out the individual effects of collinear variables on the response
- Collinearity reduces the accuracy of the estimates of the regression coefficients
- Power of the Hypothesis testing is reduced by collinearity



# Verifying Collinearity

- Ways of checking collinearity
  - Look the correlation matrix
  - Calculate the VIF (variance inflation factor)

# Verifying Collinearity

## VIF (variance inflation factor)

- Ratio of variance of  $\beta_j$  when fitting the full model divided by the variance of  $\beta_j$  if fits on its on
- Smallest possible value is 1
- A *VIF* that exceeds 5 or 10 indicates a problematic amount



# Verifying Collinearity

- What to do if a predictor has a high VIF?
  - Drop the problematic variable from the regression
  - Combine Collinear variables together into a single predictor



# Verifying Collinearity

- In R, calculating the correlation matrix is done by the following command

```
#we first can check the correlation matrix of the predictors  
cor(cereal[,c(3,4,5,6,7,8,10)])
```

```
> cor(cereal[,c(3,4,5,6,7,8,10)])
```

	protein	fat	sodium	fibre	carbo	sugars	potassium
protein	1.0000000	0.4112661	0.5727222	0.8096397	0.54709029	0.18484845	0.8417540
fat	0.4112661	1.0000000	0.2595606	0.2260715	0.18285220	0.41567397	0.3232754
sodium	0.5727222	0.2595606	1.0000000	0.4954831	0.42356172	0.21124365	0.5566426
fibre	0.8096397	0.2260715	0.4954831	1.0000000	0.20307489	0.14891577	0.9638662
carbo	0.5470903	0.1828522	0.4235617	0.2030749	1.00000000	-0.04082599	0.2420485
sugars	0.1848484	0.4156740	0.2112437	0.1489158	-0.04082599	1.00000000	0.2718335
potassium	0.8417540	0.3232754	0.5566426	0.9638662	0.24204848	0.27183347	1.0000000




# Verifying Collinearity

- The **DAAG** library provides a function to calculate the **VIF**

```
vif(multi.fit.all) # variance inflation factors  
sqrt(vif(multi.fit.all)) > 2 # problem? cutoff is 5 or 10
```

Problematic variable If  
the square of the VIF is  
greater than 2



- 
- Introduction
  - Multiple Linear Regression
  - Interaction Terms
  - Dummy variables
  - Verifying collinearity
  - Using many predictors for the fit

# Using many predictors for the fit

- Including all predictors **almost never** results in a good fit
- However, if the  $p$  is considerably small ( $\leq 10$ ) the full model is a good starting point
- Collinear predictors are the first to go
- Then p-values are verified for the remainder

# Using many predictors for the fit

- In R, there is no need to include all predictors in the linear model

```
#say we want to do a regression on all values but shelf, mfr and vitamins  
multi.fit.all = lm(calories~.-shelf -mfr - vitamins, data =cereal)
```



# Using many predictors for the fit

- In R, there is no need to include all predictors in the linear model

. means to choose  
all predictors

```
#say we want to do a regression on all values but shelf, mfr and vitamins  
multi.fit.all = lm(calories~.-shelf -mfr - vitamins, data =cereal)
```

Except the ones  
after the - sign

Qualitative variables are  
excluded for simplicity





# Next Episode

- Regression when the response is qualitative: Logistic Regression
- How to choose the best model among all possible choices?
- When the response is not normal: Generalized Linear Model



筑波大学

*University of Tsukuba*

# Multiple Linear Regression Model

Marco Antonio Florenzano Mollinetti<sup>1</sup>

<sup>1</sup>**University of Tsukuba, Systems Optimization Laboratory**

[mollinetti@syoun.cs.tsukuba.ac.jp](mailto:mollinetti@syoun.cs.tsukuba.ac.jp)