---
title: "NHANES Part II"
author: "[Your name(s) here]"
output: html_document

Quantitative Response, Quantitative Predictors: MULTIPLE LINEAR REGRESSION with INTERACTION

Continue with the NHANES example (NHANES-body.csv) for modeling weight as a function of arm circumference and arm length.

Variables:

Y: body weight in kg

X1: upper arm circumference in cm

X2: upper arm length in cm

In Part I, we considered the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where ϵ_i are independent $N(0, \sigma_{\epsilon})$.

The Interaction Model

Consider the bivariate model with interaction:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$, where ϵ_i are independent $N(0, \sigma_{\epsilon})$.

The interaction term allows the slope with respect to one predictor to change for values of the second predictor.

A. Interpretation in the Interaction Model

β_1 and β_2 do not have the same interpretation as in the bivariate linear model without interaction (Part I). Let’s investigate why, and what that interpretation is...

There are really two methods of discovering the meaning of the parameters. They both lead to the same answer/interpretation, but one or the other may make more sense to you, so I present them both below.

Method 1: Using Calculus

1. Take the derivative of the interaction model with respect to X_1.

2. Interpret this quantity in context (as you would any derivative).

3. Given the interpretation in #2, what is the change in Y (on average) when X_1 increases by 1 unit, holding X_2 constant?

4. Take the derivative of the interaction model with respect to X2.

5. Interpret this quantity in context (as you would any derivative).

6. Given the interpretation in #5, what is the change in Y (on average) when X_2 increases by 1 unit, holding X_1 constant?

Method 2: Using Algebra

1. Consider the generic model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$. Increase X_1 by 1 unit by plugging in (X_1+1) for X_1. Expand and combine like terms; call this equation Y_{new}.

2. Subtract Y from Y_{new} and cancel terms.

3. So what is the change in Y (on average) when X_1 increases by 1 unit, holding X_2 constant?

4. And similarly, the change in Y (on average) when X_2 increases by 1 unit, holding X_1 constant?

B. Fitting the Model and Performing Inference

Load data set (call it `nhanes`) and the necessary packages here:
```{r include=FALSE}
library(mosaic); library(readr); library(ggformula)
nhanes <- read.csv("~/Documents/DATA-231 F2021/Data/NHANES-body.csv") #replace this line with YOUR read.csv code
```

1. Fit the model using the `*` symbol to denote interaction.
```{r}
arm.inter <- lm(Weight~Arm.Circumference*Upper.Arm.Length, data=nhanes)
summary(arm.inter)
```

2. Compare and contrast the model summary of the no-interaction model (in NHANES Part I) with the summary of the interaction model above.

3. Write the estimated regression function.

4. Use `plot(model)` to determine if the assumptions are valid. Do you still have any concerns? Do the model assumptions seem to be met better than they were in Part I? Are there any additional remedial measures (transformations) that you feel are necessary? (Don't worry about actually taking any of those actions right now, just discuss any problems you see.)

5. Conduct an appropriate statistical inference to determine if the interaction term should be included in the model. (Don’t worry about checking conditions, since you discussed them in #4).

6. Notice that because of the interaction term, the relationship between Weight and either of the two predictor variables is more complicated:

  a. What is the relationship between weight and arm length, for those people with arm circumference of 30 cm?

  b. What is the relationship between weight and arm length, for those people with arm circumference of 40 cm?

7. The first row of this dataset is an individual with upper arm length of 35.6cm and upper arm circumference of 40.5cm.

 a. The code below will calculate the predicted/estimated weight for that person using the linear model without interaction (from NHANES Part I) and the interaction model. Notice the difference in predictions between the two models.

```{r}
arm.lm <- lm(Weight~Arm.Circumference+Upper.Arm.Length, data=nhanes) #fit the no-interaction model again
arm.lm$fitted.values[1] #predicted weight of Person 1 based on no-interaction model
arm.inter$fitted.values[1] #predicted weight of Person 1 based on interaction model
```

  b. Now find and compare the residual for that person using the no-interaction model and the interaction model.

D. Comparison of Models
----

1. If you felt a transformation should have been taken based on the residual analysis, take that transformation now. Show the model summary and the residual plots for this new model.

2. Which model (bivariate linear without interaction or bivariate linear with interaction, or some transformed model you've discovered) do you think is "best"? Explain why, including your criteria for "best".