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

Quantitative Response, Quantitative Predictors: MULTIPLE LINEAR REGRESSION with INDICATOR
VARIABLES
-----------------------------------------

Continue with the NHANES example (NHANES-body.csv) for modeling weight as a function of arm
circumference and arm length.  In the original (no-interaction) model, arm circumference was
the more significant predictor of weight, so for now we will use just arm circumference,
along with an indicator variable for sex.

**Variables:**

Y:  body weight in kg

X1: upper arm circumference in cm

X2: sex (0=male, 1=female)

A.  MODEL: No-Interaction
--------------------------
Suppose we consider the bivariate model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +
\epsilon$, where $\epsilon_i$ are independent $N(0, \sigma_{\epsilon})$.

1. Suppose that $X_2 = 0$ (males).  What is the resulting model?


2. Suppose that $X_2 = 1$ (females).  What is the resulting model?


3. Under what conditions will the model be the same for males and females?


4. What is the change in $Y$ (on average) when $X_1$ increases by 1 unit for males?


5. What is the change in $Y$ (on average) when $X_1$ increases by 1 unit for females?


**Lesson**: Including an indicator variable (but no interaction term) is allowing the groups
(male vs. female, in this case) to have different *intercepts*.  That is, an indicator
variable will lead to different, but *parallel*, regression lines.


B. MODEL: Interaction
-----------------------------

Suppose we 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 $\epsilon_i$ are independent $N(0,
\sigma_{\epsilon})$.

1. Suppose that $X_2 = 0$ (males).  What is the resulting model?


2. Suppose that $X_2 = 1$ (females).  What is the resulting model?


3. Under what conditions will the model be the same for males and females?
```

~~~~
4. What is the change in $Y$ (on average) when $X_1$ increases by 1 unit for males?


5. What is the change in $Y$ (on average) when $X_1$ increases by 1 unit for females?


**Lesson**: Including an indicator variable *with* an interaction term is allowing the groups
(male vs. female, in this case) to have different *slopes and intercepts*.  That is, an
indicator variable with interaction will lead to completely different (not parallel)
regression lines.


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



Make scatterplots of weight (Y) vs. arm circumference (X) for men and women separately.  You
can do this on separate panels, using
```{r fig.width=7, fig.length=5}
gf_point(Weight~Arm.Circumference | Gender, data=nhanes)
```

1. Describe the main features of the relationship between arm circumference and weight for
males and for females.



It may help to look also at all points on the same grid, with different colors AND different
shapes for the genders.
```{r}
gf_point(Weight~Arm.Circumference, shape=~as.factor(Gender), color= ~as.factor(Gender),
data=nhanes)
```

Because most of the points lie on top of each other, it's hard to see what's happening here.
You can change the opacity (i.e., make points more see-through) with the `alpha` argument.

```{r}
gf_point(Weight~Arm.Circumference, shape=~as.factor(Gender), color= ~as.factor(Gender),
alpha=0.5, data=nhanes)
```

2. Does it appear that the estimated regression function for males will be the same as for
females?  How would they be alike?  How would they differ?


D. No-Interaction Model
--------------------------------
We already know that the relationship between weight and arm circumference gives us a model
with non-constant variance. To (pre-emptively) deal with this problem, **use Y= log(Weight)
as your response variable.**

1. Fit the estimated no-interaction regression model (with arm circumference and gender as
predictors) from Part A.
```{r}

```

2. Write the estimated regression function (using variable names, of course).
~~~~

3. Write the estimated regression function *for males*.

4. Write the estimated regression function *for females*.

5. Use residual plots to determine if the regression assumptions are valid.

6. Conduct appropriate statistical inference to determine if the intercept parameter is the same for males and females.

7. State the hypotheses for the ANOVA F-test for the entire model, and report the conclusion of the test.

Visualization:

8. Plot the males with their fitted regression line and the females with their fitted regression line on the same plot. You can use whatever plotting function you want here, but make sure that: everything's on the same graph (not different panels); there are different colors and different shapes for the 2 genders; there is a useful legend.

```{r}

```

9. Do you think that a separate intercept term is necessary here?

E. Interaction Model

Again, make sure you **use Y= log(Weight) as your response variable.**

1. Fit the estimated regression model with interaction from Part B.

```{r}

```

2. Write the estimated regression function (using variable names, of course).

3. Write the estimated regression function *for males*.

4. Write the estimated regression function *for females*.

5. Interpret the value 0.04492. That is, what does this value tell us about the relationship between arm circumference and log(weight)? Then tell me what we can say about the relationship between arm circumference and weight.

6. Use residual plots to determine if the regression assumptions are valid.

7. Conduct appropriate statistical inference to determine if the intercept parameter is the same for males and females (using this model).

8. Conduct appropriate statistical inference to determine if the slope parameter is the same for males and females.

9. State the hypotheses for the ANOVA F-test for the entire model, and report the conclusion of the test.

Visualization:

10. Plot the males with their fitted regression line and the females with their fitted regression line on the same plot. You can use whatever plotting function you want here, but make sure that: everything's on the same graph (not different panels); there are different colors and different shapes for the 2 genders; there is a useful legend.

~~~
```{r fig.width=5, fig.length=5}

```
~~~

11. Do you think that a separate slope term is necessary here?

F. Alternative Approach

Rather than using a multiple regression model with an indicator variable for sex, we could have done two separate simple linear regressions (with arm circumference as the X variable), one for males and one for females.

There are lots of ways to create a "females" data set and a "males" data set, but the function `subset` or `filter` could be useful.

Fit the two models, then compare and contrast the two approaches (bivariate model with interaction vs. two separate models). What are the advantages/disadvantages of each? Can you imagine situations where someone might prefer one approach over the other?