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

Quantitative Response, Quantitative Predictors: MULTIPLE LINEAR REGRESSION

ORIGINAL SOURCE: National Health and Nutrition Examination Survey, Centers for Disease Control and Prevention, http://www.cdc.gov/nchs/nhanes.htm.

Usually, body weight is determined by weighing on a scale. But what if a scale is unavailable? Could body weight be estimated using other, more easily measured, body measurements? Suppose we only have a tape measure available, so we can only consider the body measurements obtained using a tape measure. We would like to model weight as a linear (we hope!) function of length and circumference measurements, following "Occam's razor" or the "law of parsimony", which says "entities should not be multiplied beyond necessity" or "Of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred."

That is, we would like to have an "estimating equation" for body weight that provides the "best" estimate *with the fewest variables*.

We will be using data from NHANES 2009-2010 data to model body weight as a linear function of arm circumference and arm length. We will restrict attention to only those individuals who are between 18 and 24 years old (inclusive). That is, we will restrict attention to "young adults".

Variables:

Y: body weight in kg

X1: upper arm circumference in cm

X2: upper arm length in cm

Data set: on Moodle as "NHANES-body.csv". Download the file, then load it into the Console, then cut-and-paste the `read.csv` code into the chunk below.

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
```

A. Investigating the Data
---

1. You can read about the NHANES survey in the NHANES brochure (on the Moodle page). From the information provided, do you feel the ever-present assumptions of representativeness and independence are satisfied with this data? Why or why not?

2. **Univariate Distributions:** Using graphical and descriptive statistics, describe the main features of the distribution of each variable, noting any unusual features.

3. **Bivariate Relationships:** We'd like to look at the relationship between all three of these variables. Of course, we could just make three xyplots, but it might save a little time if we construct a scatterplot matrix. This is a graph that makes scatterplots for all possible pairs of variables.

a. Make the scatterplot matrix using the `pairs()` function. To specify only these three variables (instead of all variables in the data set), use

```{r fig.height=5,fig.width=5}
pairs(~Weight+Upper.Arm.Length+Arm.Circumference, data=nhanes)
```

b. For each pair of variables, describe the main features of the relationship, noting any unusual features.

B. Multiple Regression Model

The simplest multiple linear regression model with two independent variables is:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where ϵ_i are independent $N(0, \sigma_{\epsilon})$.

1. How many unknown parameters are in this model? (Note that these parameters are the ones that will be estimated by the least-squares method.)

2. The formulas for the least-squares estimates are similar, but more extensive, than those for the simple linear regression. Of course, we will use R to calculate these estimates for us, using our old friend `lm()`. The only change is we now indicate more than one predictor variable with a "+" sign in the formula argument:

```{r}
arm.lm <- lm(Weight~Arm.Circumference+Upper.Arm.Length, data=nhanes)
summary(arm.lm)
```

3. The fitted model is:

4. Interpret the estimates $\hat \beta_1$ and $\hat \beta_2$ in context.

C. Diagnostics and Residual Analysis

Use the usual `plot(model)` to investigate the residual plots.

1. Verify the validity of the model assumptions using the residual plots.

```{r}

```

2. What, if any, remedial measures (transformations or otherwise) are indicated by the residual analysis and diagnostics?

3. Carry out any "fixes" you feel are necessary (but don't add any new terms to the model yet!) and re-run the multiple linear regression model. Verify the validity of the new model using residual plots. Did you fix the problem(s) to your satisfaction?

D. Inferences About β Parameters

In this section, use your "final" model from Part C, #3.

1. Explain why it is not appropriate in this setting to make an inference about β_0.

2. Construct a 90% confidence interval for β_1 and interpret the interval in context.

3. Conduct a test of significance for β_2.

4. Conduct the test of significance **for the entire model** using the ANOVA table.

5. Interpret R^2 in context.

E. Practical Significance/Effect Size

When conducting tests of statistical inference, we should be careful to not conflate statistical significance with practical significance. Statistical significance comes from p-values and answers the question: Is the difference between Ho and the data larger than we would expect just by random chance? Practical significance answers the question: Do we care?

Harvard statistician A. P. Dempster phrased this as "IS questions" and "IT questions": "IS there an effect?" can be answered by p-values. "How big is IT (the effect)?" is a question about effect size. **It's important to realize that the p-value can not tell you how strong the effect (in this case, the association/correlation) is, nor how well we can estimate it.** To do this, we need confidence intervals!

In this case, both β_1 and β_2 are *statistically* significant, but are they *practically* significant? That is, are the effects of these two variables on Weight large enough to make a substantial difference, or to be "useful"? Which one is more practically significant? Discuss with your Table.

Of course, reasonable people can disagree about what a "large effect" is, what "useful" means, and what is "practically significant". This is part of the reason that these issues are traditionally ignored in favor of the objective certainty of a (mathematical) p-value. But once again, we must embrace the nuance and potential disagreement that these questions raise. Many things are statistically significant but not practically significant! And *you* may not know whether something is practically significant in a certain situation. This is why interdisciplinarity is so important! Suppose I find a statistically significant relationship between air temperature (x) and the length of butterfly wings (y), with a slope between the two of -0.24 mm (95% CI: -0.41 mm, -0.07 mm). Is this practically significant? I have no idea! I don't know anything about butterflies! But I *should* be working with a butterfly specialist, and it's *their* job to decide if this is a practically significant effect.