~~~
---
title: "Simple Linear Regression: Vitruvius I"
author: "[Your names here]"
output: html_document
---

```{r, include = F}
# load packages we typically use for this class.
library(mosaic)
library(ggformula)
library(Stat2Data)
library(tidyverse)
~~~

A. Research Question
------------------------------------
Marcus Vitruvius, an architect in Rome, wrote in De Architectura:

“For if we measure the distance from the soles of the feet to the top of the head, and then
apply that measure to the outstretched arms, the breadth will be found to be the same as the
height, as in the case of plane surfaces which are completely square.”
(Marcus Vitruvius, De Architectura, Book III, Chapter 1, p 3)

Using Vitruvius’ claim, this is our research question: Do people have heights (X) that are in
a linear relationship with their wingspan (Y)?

(We're probably specifically interested in whether the slope of this linear relationship is
equal to 1, but for now let's just see if there's *any* linear relationship...)

B. Data Definition and Collection
------------------------------------
We will use a convenience sample to assess the validity of Vitruvius’ claim.  Our data are
the heights and wingspans of the students from Data 231 last fall.

1. Clearly describe an individual in our study.

2. the response variable is:

the explanatory variable is:

3. Clearly describe how we collected/measured the data, including the measuring device used.
(The idea here is to be specific enough that another person could repeat your study exactly.)

[*I have done this for you!*]
When measuring the wingspan measure longest finger to longest finger, with arms outstretched.
When measuring height, measure from soles of feet to crown of head, without shoes. Each
person was measured separately by a different measurer.  Units are centimeters.

Upload data set from the Google Sheet.  To do this, we'll use the package 'gsheet', which is
NOT installed on Wooster's server.  Install the package now, then run the code below.
```{r}
library(gsheet)
Woo.Data <- gsheet2tbl('docs.google.com/spreadsheets/d/1NbLdAa7urCG4_AsSP-
S8aKnHX_l3LGeB6YInFwaNcjs/edit?usp=sharing')
```

C. Exploratory Data Analysis
------------------------------------
1. Distribution of Response Variable.
```
```

~~~
Construct an appropriate visual display of the distribution of the response variable,
describe its main features, and note anything unusual.


2. Distribution of Explanatory Variable.
Construct an appropriate visual display of the distribution of the explanatory variable,
describe its main features, and note anything unusual.


3. Construct a scatterplot and describe the form, direction, and strength of the relationship
between the two variables.


4. Calculate the correlation between the two variables.


D. Model Fitting
----------------------------
1. Fit the regression line.

We do this with the standard `function(Y~X, data=data_name)` format we've been using. The
function is `lm`, which stands for "linear model". Let's call the model `vitruvius` so we can
refer back to it later.
```{r}
#vitruvius <- lm(Y~X, data=data_name)

```


2. Interpret the slope in the context of the problem. You can see the slope and intercept by
typing `vitruvius`.


3. Interpret the intercept in the context of the problem.  Does the intercept make sense
here?  That is, does it have a practical interpretation?



### Regression Summary:

Notice that if you type just
```{r}
vitruvius
```
you don't get much information.  But `vitruvius` is an **object** that *contains* lots of
information within it.  For something more informative, type
```{r}
summary(vitruvius)
```

You can also ask for the residuals (for each observation) and the fitted $\hat y$ values (the
model's predicted $y$ for each observation) with
```{r}
residuals(vitruvius)
fitted(vitruvius)
```


E. Graphing the Least Squares Line and Assessing the Fit
---------------------------------------------------------

1.Add the fitted line to your scatterplot from C.3

You do this by adding `geom_abline`, then specifying the slope and intercept.
```{r fig.width=4, fig.height=4}
#gf_point(Y~X, data = data_name) +
~~~

~~~~
geom_abline(slope=your_slope_value,intercept=your_intercept_value)
```


### Assess

A simple linear regression model can be useful for three things:
  - as a summary of the relationship between two numerical variables;
  - as a guide for predicting a future value of $y$ based on a particular value of $x$;
  - as a basis for making a decision about the statistical significance of the slope or
correlation.

The first bullet point is part of **exploratory data analysis**; i.e., using data to
summarize.  The only condition required is that the relationship between the two variables
must be linear.  The last two bullet points fall under the heading of **statistical
inference**; i.e., using sample data to make a claim about the population.  In this case, we
require additional conditions about the residuals to be true.

2. Based only on the graph with fitted line in E.1, does it appear that the line is good
summary of the relationship between these two variables?



3. Do you think the conditions for inference are satisfied?

Use the code below to make the appropriate graphs to answer this question.  (See Example 1.5
if you need a guide for this question.)  (Note that `plot(vitruvius)` gives you more plots
than you need!  Just ignore the extraneous ones for now.)
```{r fig.width=8, fig.height=8}
par(mfrow=c(2,2))
plot(vitruvius)
```



4. Based on your plots in E.1 and 3, are any transformations necessary?  If so, make those
transformations and repeat your analysis.



F. Inference
----------------------------
For significance tests, be sure to state the hypotheses and state your conclusion in context.
You do not need to check the conditions because you already did that in Section E.

1. Test the claim that the correlation coefficient is not 0.


2. ANOVA table:
```{r}
#the code is anova(model.lm), where model.lm is the model you fit in part D
```



3. Using the ANOVA table, test if the model explains a significant amount of the variability
in the response variable.


4. Calculate and interpret a 95% CI for the slope of the line relating Wingspan & Height.
```{r}
#The code to do this is:
#confint(vitruvius)
#95% is the default conf level, but you can change it by adding the argument 'level=...'
```

5. We want to calculate a 90% interval for the **mean Wingspan** of people whose Height is
~~~~

170 cm (67").

~~~
```{r}
#Code for confidence interval:
#predict.lm(model.lm, data.frame("Height"=170), interval="confidence", level=0.9)
#Code for prediction interval:
#predict.lm(model.lm, data.frame("Height"=170), interval="prediction", level=0.9)
```
~~~

a. Do we want a prediction interval or a confidence interval?

b. Calculate the interval.

c. Interpret the interval.

6. We want to calculate 90% interval for a **person's Wingspan** when their Height is 170 cm (67").

a. Do we want a prediction interval or a confidence interval?

b. Will this interval be wider or narrower than the interval in #5b?

c. Calculate the interval.

d. Interpret the interval.

G. Conclusion
-------------------

1. Based on your analysis, Do people have heights that are in a linear relationship with their wingspan?

2. Do you have any concerns about the generalizability of this data set?

3. Do you concur with Vitruvius’ claim about the *specific relationship* between height and wingspan?