STAT 113: HW5

Do Not Circulate!

Last Revised March 25, 2022

Specific Learning Objectives:

B4: Using Correlation and Regression to Relate One Quantitative Variable to Another

B5: Recognizing and Diagnosing the Appropriateness of Regression Models

D1: (Lab5) Producing Descriptive Summaries and Visualizations

D2: (Lab5) Techniques for Transparent and Reproducible Results

Problems:

1. (B4) **Genetic Diversity and Distance from Africa** It is hypothesized that humans originated in East Africa, and migrated from there. We compute a measure of genetic diversity for different populations, and the geographic distance of each population from East Africa (Addis Ababa, Ethiopia), as one would travel over the surface of the earth by land (migration long ago is thought to have happened by land). The relationship between these two variables is shown in Figure 1 and the data is available `GeneticDiversity` from the `Lock5Data` R package.

   (a) Describe the relationship between genetic diversity and distance from East Africa. Does there appear to be an association? If so, it is positive or negative? Strong or weak? Linear or nonlinear?

   (b) Which of the following values gives the correlation between these two variables: $r = -1.22$, $r = -0.83$, $r = -0.14$, $r = 0.14$, $r = 0.83$, or $r = 1.22$?

   (c) On which continent is the population with the lowest genetic diversity?

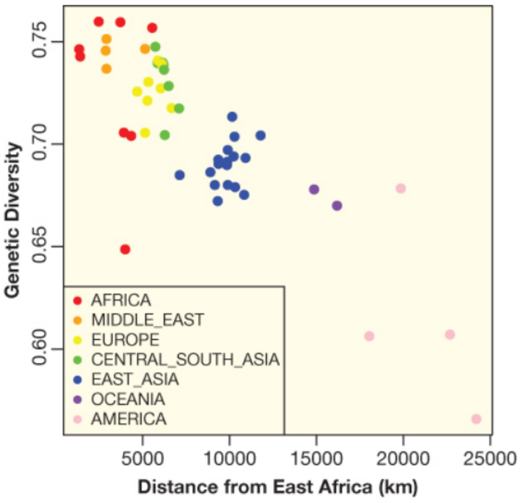

Figure 1: Genetic diversity of populations by distance from East Africa

On which continent is the population that is farthest from East Africa (by land)?

(d) Populations with greater genetic diversity are thought to be better able to adapt to changing environments, because more genetic diversity provides more opportunities for natural selection. Based only on this information and Figure 2.57, do populations closer or farther from East Africa appear to be better suited to adapt to change?

2. (B4) **Football and Cognitive Percentile** A study examined several variables on collegiate football players, including the variable `Years`, which is number of years playing football, and the variable `Cognition`, which represents the percentile at which a person scores on a standardized battery of cognitive tasks.

The correlation between these two variables calculated on the dataset collected by the researchers is -0.366. The regression linec for predicting `Cognition` from `Years` is:

$$\widehat{\texttt{Cognition}} = 102 - 3.34 \cdot \texttt{Years}$$

(a) Predict the cognitive percentile for someone who has played football for 8 years and for someone who has played football for 14 years.

(b) Interpret the slope in terms of football and cognitive percentile. Be sure to ground your interpretation in the concrete context.

(c) All the participants had played between 7 and 18 years of football. Is it reasonable to treat the intercept as telling us something meaningful about football players in itself? Why or why not?

3. (B4) **Is the Honeybee Population Shrinking?** The Honeybee dataset contains data collected from the USDA on the estimated number of honeybee colonies (in thousands) for the years 1995 through 2012. The regression line to predict number of colonies (in thousands) from year (in calendar year) that minimizes the sum of squared residuals for this dataset is

$$\widehat{\texttt{Colonies}} = 19,291.511 - 8.358 \cdot \texttt{Year}$$

(a) Interpret the slope of the line in context.

(b) Often researchers will adjust a year explanatory variable such that it represents years since the first year data were collected. Why might they do this? (Hint: Consider interpreting the y-intercept in this regression line.)

(c) Predict the bee population in 2100. Is this prediction appropriate? Why or why not?

4. (B5) The plots in Figure 2 are **residual diagnostics** from a linear regression model reported in a paper entitled, "Skull dimensions in relation to body size in non-human mammals", published in 2000 by W. Tecumseh Fitch. Each data point corresponds to a **non-human mammal species**.

The **explanatory variable** in the model is the **weight of the mammal's body in kilograms (kg)**.

The response variable in the model is the **length of the mammal's skull in centimeters (cm)**.

The plot on the left shows the **fitted values** from the regression model on the x-axis and the **residuals** on the y-axis. The solid blue line is a **rolling average of the residuals**.

The plot on the right is a **histogram of the residuals** from the regression model, along with an **estimated density curve**.

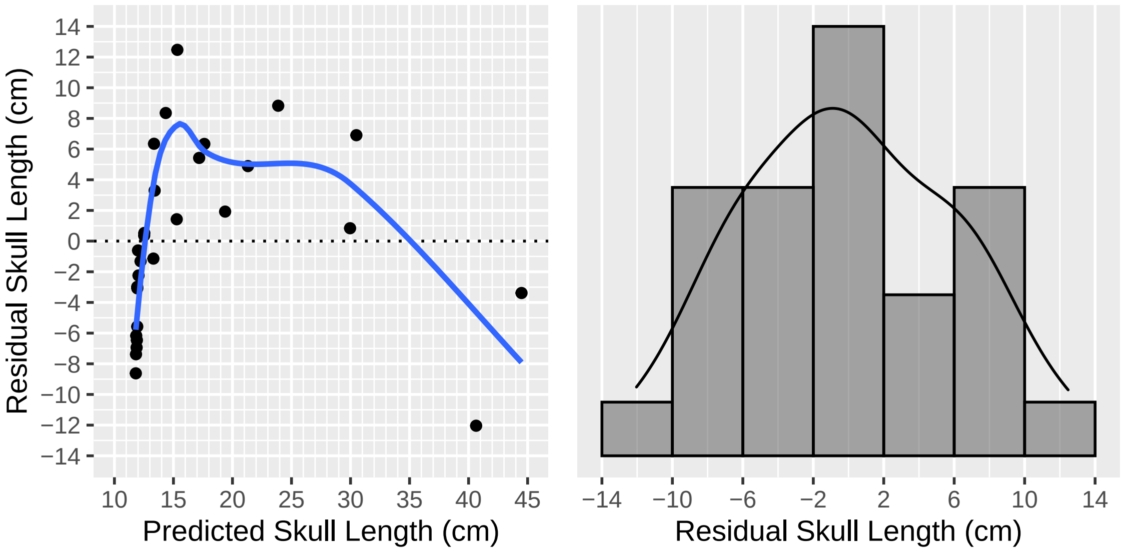

Figure 2: Residual plots for a linear regression model predicting Skull Length of non-numan mammal species from their body weight in kilograms

(a) Do you have any concerns about the validity of the linear model based on the graph on the left? Explain.

(b) Do you have any concerns about the validity of the linear model based on the graph on the right? Explain.