MATH108: Elementary Statistics

Spring 2014

In-Class Activity 01: Exploratory Data Analysis & Regression

Assignment is DUE as indicated on the course schedule. This assignment is designed to be completed in class.

*This is a **group assignment**. Work with 3-5 classmates, and submit as a group on Gradescope.*

Notes

Your final submission must be readable. It is your responsibility to write up your answers in a way that is easy to read and follow.

All group members are expected to contribute to all parts of this assignment.

Part 1

For this assignment you will work with the 5k.csv data under the "Demos" tab on the course website.

Open the dataset and familiarize yourself with it. What does one observation in this dataset represent? What are the variables and what are their types? Record your answers to these questions in a document.

Given this data, our main questions of interest are:

1) Are "age" and "net_sec" associated?
2) Are "over40" and "net_sec" associated?

To answer these questions, start with some exploratory data analysis. Generate the appropriate summary visualizations for:

- age
- over40
- net_sec

- age vs net_sec
- over40 vs net_sec

Put your visualizations into your document. Alongside each, describe what it shows about the data.

Part 2

Next, we'll look at a few regressions to further investigate our two questions.

We will start with question #1. Here are summary statistics for age and net_sec:

n <int>	mean_age <dbl>	s_age <dbl>	mean_net_sec <dbl>	s_net_sec <dbl>
2515	38.04175	12.74367	2375.577	586.4648

Use these statistics to find the coefficient estimates for the regression model:

$$net\_sec = \beta_0 + \beta_1 age$$

Based on your model:

- What is the predicted net_sec for a runner of age 0?
- What is the predicted change in net_sec given a one-year increase in age?

Now, we will look at question #2. Here are summary statistics for over40 and net_sec:

n <int>	mean_over40 <dbl>	s_over40 <dbl>	mean_net_sec <dbl>	s_net_sec <dbl>
2515	0.3829026	0.4861915	2375.577	586.4648

Use these statistics to find the coefficient estimates for the regression model:

$$net\_sec = \beta_0 + \beta_1 over40$$

Based on your model:

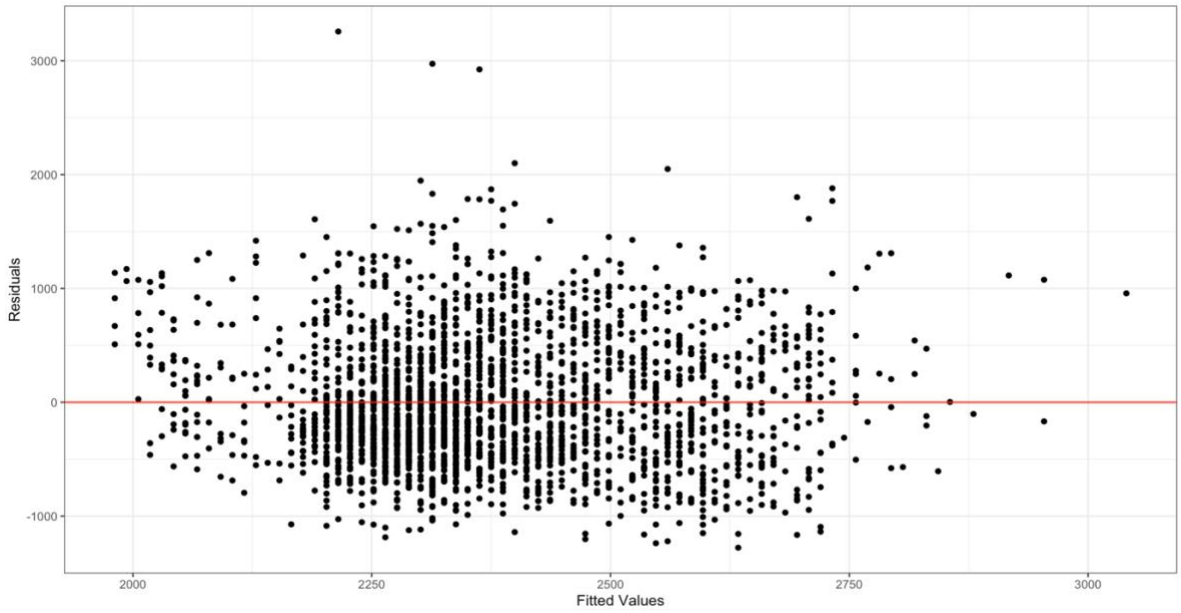
- What is the predicted net_sec for a runner under age 40?
- What is the predicted net_sec for a runner over age 40?

Part 3

Below are the R^2 statistics and residual plots for each regression. Based on these do you think one is a better fit than the other? Why?

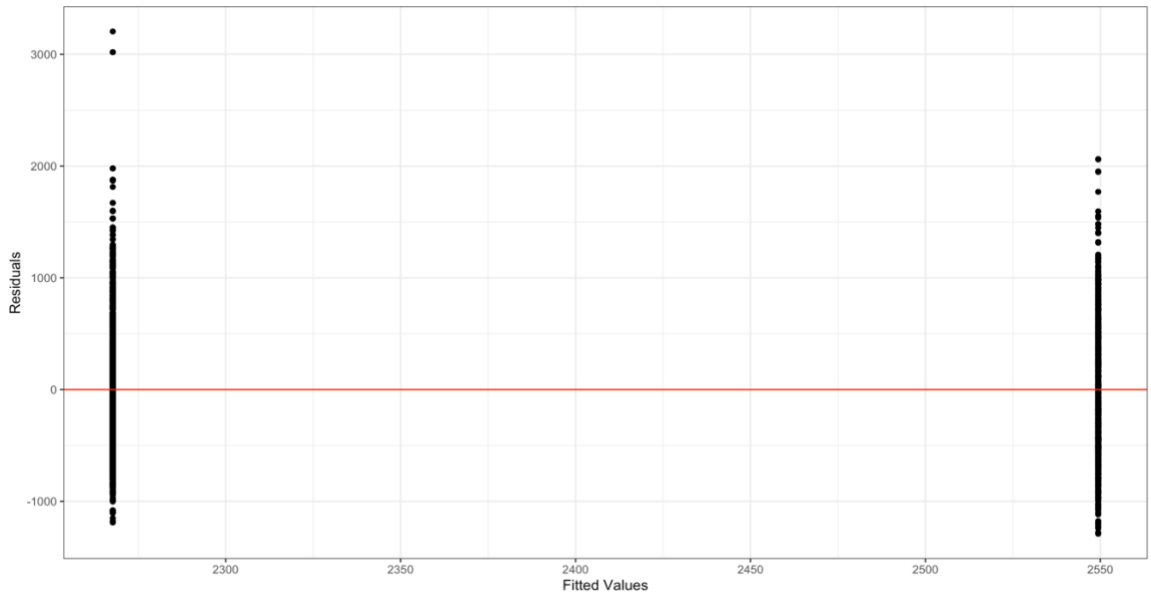
$$net\_sec = \beta_0 + \beta_1 age$$

$R^2 = 0.072$
Residual plot:



$$net\_sec = \beta_0 + \beta_1 over40$$

$R^2 = 0.055$
Residual plot:



Submission

Submit on Gradescope. Remember to tag your groupmates!

Rubric

	Missing / Not Complete (0)	Approaching (2)	Meets (4)	Exceeds (5)
Readability	Assignment is unreadable or not submitted.	Assignment includes formatting, but significant improvements could be made. For example, clear labeling of problems and subparts, proofreading.	Assignment includes formatting, but minor improvements could be made. For example, clear labeling of problems and subparts, proofreading.	Assignment is well formatted and easy to read. Text has been proofread.
Completeness	Less than half of assignment is attempted.	Roughly half of assignment has been attempted. On the problems that have been completed, effort is evident. OR All of the assignment has been attempted, but effort is not evident in many parts.	At least 80% of assignment has been attempted. On the problems that have been completed, effort is evident. OR All of the assignment has been attempted, but effort is not evident a few parts.	All of the assignment has been attempted, and effort evident throughout.
Correctness	All answers are incorrect or missing.	Of the complete problems, at least half have been approached and completed correctly.	Of the complete problems, at least 80% have been approached and completed correctly.	All complete problems are approached and completed correctly.