KENAN FELLOWS PROGRAM
FOR TEACHER LEADERSHIP

Applying Linear Regression to Marathon Data

AP Statistics
Mrs. Angela Stabler

Description

This lesson/project is a culminating activity designed to reinforce the concepts learned about linear regression, as well as demonstrate the mastery of learning these concepts. The teacher will have covered all notes regarding linear regression and the students will have had some practice with these concepts.

Packet Contents

Introduction

Linear regression is used to model the relationship between two variables and to explain the strength of the relationship. It can be used to forecast or predict outcomes based on the overall model for the observed data set. Linear regression is used and applied in a variety of concentrations including biological, healthcare and pharmaceutical data, athletic data, economic data, etc. In this lesson/project the students will be determining linear regression models for marathon data. Their goal is to predict a runner's finishing time for a race based on the runner's intermediate splits, such as 5k, 10k, and half marathon split times. A split time is the time it takes for a runner to run a smaller portion of the race, ie a 5k split would be how long it takes for a runner to run the first 5 kilometers. The data sets that will be used include Boston Marathon runner data, the City of Oaks Marathon runner data, and 5k runner data from Enloe High School cross country runners. This lesson is to be used as an assessment after all material for this unit has been taught.

Curriculum Alignment

Common Core Standard Course of Study NC- These also correspond and fulfill the Advanced Placement Standards.

Summarize, represent, and interpret data on two categorical and quantitative variables.

S-ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models.

b. Informally assess the fit of a function by plotting and analyzing residuals.

c. Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models.

S-ID.7 Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

S-ID.8 Compute (using technology) and interpret the correlation coefficient of a linear fit.

S-ID 9 Distinguish between correlation and causation.

Objectives

Students Will Be Able To:

- Create a scatter plot from bi-variate data using a data set and Excel Software.
- Make a best fit regression line for the data.
- Interpret slope, y-intercept, correlation coefficient and r^2 in context.
- Determine outliers and influential plots.
- Use computer output data to determine the best fit regression line.
- Use computer output data to interpret the correlation coefficient and coefficient of determination.
- Find residual for a given data point.
- Create and interpret, in context, a residual plot.
- Use computer output data to determine the best model that fits the data.
- Recognize how the slope, y-intercept, standard deviation of the residuals, and r^2 are influenced by extreme observations.
- Predict runners' times using models found.

Time & Location

Each class period is 86 minutes. This lesson is expected to take approximately 2 class periods.

Teacher Materials

- A computer lab with 32 computers and access to Excel
- Master copy of the assignment and assessment
- Notes/assignment handout- 32 copies (if teacher chooses to)

Student Materials

- A computer with Excel
- The Practice of Statistics (5th edition) textbook (or any textbook to aid them)
- Notes (given by teacher throughout the unit)

Safety

- Ensure students are aware of rules of safe computer use and plagiarism

Student Prior Knowledge

Student need to have prior knowledge of:

- Creating a scatter plot using bi-variate data
- Making a best fit regression line.
- Interpreting slope, y-intercept, correlation coefficient and r2 in context.
- Determining outliers and influential plots.
- Finding residual for a given data point.
- Creating and interpreting in context a residual plot.
- Recognizing how the slope, y-intercept, standard deviation of the residuals, and r2 are influenced by extreme observations.
- Proper graphing techniques and elements

Teacher Preparations

The teacher needs to provide access to the following for each student

- **Regression Runner Data** (the student will need to download the doc-it will not work through google sheets)
- **Excel How To: Regression Project**
- **Linear Regression Practice on Marathon Data**
- **Excel How To: Regression Project**
- **Linear Regression Runner Data Assessment (Individual)**

Activities

At the start of class the students will find and log into a computer. The teacher will instruct the students to open the **Excel How To: Regression Project** document and the **Linear Regression Practice on Marathon Data** document. The **Linear Regression Practice on Marathon Data** is the guided assignment. The guided assignment should take the students between 45-60 minutes. The teacher will instruct the students to:

- Open the Excel How To: Regression Project document on their computers (I have it posted to my website and my Google Classroom).
- Turn on the Data Analysis ToolPack in Excel (using the **Excel How To: Regression Project** guide)
- Open the link to the data in Google drive (if the teacher is using their own Google Drive) or open the data from the teacher's website.
- Begin working through the **Linear Regression Practice on Marathon Data** Assignment

As the students are working on the assignment, the teacher will monitor and assist students as needed. The students will finish the assignment at different paces.
Please see the *Linear Regression Practice Answer Key* for a model of appropriate answers.

Assessment

Once each student has finished the teacher will instruct the students to open the **Linear Regression Runner Data Assessment (Individual)**assessment. The students will complete and should submit this as evidence of their skill mastery.

For this assessment the students will perform the same analysis on the City of Oaks Race data (the students will use all of this data), located on sheet 2 of the Regression Runner Data spreadsheet. The students will develop the times for a half marathon (using each runner's 10k time) and marathon (using each runner's half marathon time) for a sample of runners given. See the prediction data sets on next page for what the students will use for predictions. They will only perform predictions for one of the three data sets. They are to discuss how extrapolation effects the results. They will then make a poster/presentation to report their results and findings. This assessment is to be done individually without assistance. It should be submitted as a poster.

See below for the **Predictions Data Sets** that the students will need to complete predictions for this project. There are three different sets to allow the teacher to distribute to the student to minimize the amount of students copying from other students.

- See the *Linear Regression Runner Data Assessment* assignment for full details and expectations.
- Please see the *Marathon Data Regression Rubric* for the poster rubric.
- Please see the *Linear Regression Assessment Answer Key* for a model of appropriate answers.

Predictions Data Sets

Prediction Data Set 1		
Runner	10k Split (min)	Half Marathon Split (min)
1	48.8	101.52
2	55.88	112.56
3	61.57	127.23
4	105.45	240.07
5	36.38	75.87
Prediction Data Set 2		

Runner	10k Split (min)	Half Marathon Split (min)
1	38.14	77.89
2	59.11	119.42
3	62.65	126.3
4	108.24	217.36
5	35.2	71.32
Prediction Data Set 3		
Runner	10k Split (min)	Half Marathon Split (min)
1	46.99	94.67
2	56.49	114.79
3	65.84	132.96
4	112.45	226.91
5	32.28	65.46

Critical Vocabulary

- Mean- arithmetic average; add all the values of the data set and divide by the number of observations.
- Standard deviation- measures the typical distance of values in a distribution from the mean
- Slope- the amount by which *y* is predicted to change when *x* increases by one unit
- Scatterplot- plot that shows the relationship between two quantitative variables measured on the same individuals.
- Bivariate-two variable data
- Correlation, r- measures the directions and strength of the linear relationship between two quantitative variables.
- Coefficient of Determination, r^2-fraction of the variation in the values of *y* that is accounted for by the least-squares regression line of *y* on *x*.
- Residual- difference between an observed value of the response variable and the value predicted by the regression line.
- Linear Regression- an approach for modeling the relationship between a scalar dependent variable *y* and one or more explanatory variables denoted *x*.

- Least Squares Regression Line- line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the values of y for a given value of x.
- Outlier- individual value that falls outside the overall pattern of a distribution
- Influential Point-an observation that if removed would markedly change the result of the calculation
- Extrapolation- use of a regression line for prediction far outside the interval of values of the explanatory variable x used to obtain a line. Such predictions are often not accurate.

Author Information

Kenan Fellow: Angela Stabler

- Enloe High School
- Grades 9-12, AP Statistics, AP Research, Common Core Math I
- Teaching since 2012
- astabler7@gmail.com

Mentor: Dr. Richard Smith

- Email: rls@email.unc.edu
- Department of Statistics and Operations Research, University of North Carolina,
- Richard L. Smith is Mark L. Reed III Distinguished Professor of Statistics and Professor of Biostatistics in the University of North Carolina, Chapel Hill. He is also Director of the **Statistical and Applied Mathematical Sciences Institute**, a Mathematical Sciences Institute supported by the National Science Foundation.His main research interest is environmental statistics and associated areas of methodological research such as spatial statistics, time series analysis and extreme value theory. He is particularly interested in statistical aspects of climate change research, and in air pollution including its health effects. He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, an Elected Member of the International Statistical Institute, and has won the Guy Medal in Silver of the Royal Statistical Society, and the Distinguished Achievement Medal of the Section on Statistics and the Environment, American Statistical Association. In 2004 he was the J. Stuart Hunter Lecturer of The International Environmetrics Society (TIES). He is also a Chartered Statistician of the Royal Statistical Society.