PROJECT-SET
Statistics Education for Teachers

Used Car Regression

Overview of Lesson
Simple Least Squares Regression is used to model the linear relationship between two quantitative variables, X and Y, where X is the explanatory variable and Y is the response variable. This lesson focuses on necessary content for A.P. Statistics students, including using data to create an appropriate model that maximizes R^2, while satisfying the underlying assumption of linearity and transforming data to achieve linearity where necessary.

This task requires students to collect real data on two quantitative variables and conduct a regression analysis to predict the price of a car from the age of a car. Specifically, the students will use the classified ads from a local newspaper (or an online source) to collect the data on the age in years and the price of a specific car make and model (e.g., Ford Mustang or Honda Accord). The website http://www.carsforsale.com/ allows the student to select the car make and model, which will produce a list of cars in order of most recent entry to oldest entry. Some models of cars seem to devalue more quickly in the early years. As a result, the scatterplot and subsequent residual plot may indicate that the relationship between the two variables is not completely linear. The student's task is to improve the model using a transformation, such as a log X, log Y, or logX-logY. Improvements will be noted by an increase in R^2 and a residual plot that shows no pattern.

GAISE Components
This activity follows all of the four components of the statistical problem-solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) *Report.* This is a Level C activity.

Common Core State Standards for Mathematical Practice
1. Model with mathematics.
2. Make sense of problems and persevere in solving them.
3. Reason abstractly and quantitatively
4. Use appropriate tools strategically.
5. Attend to precision.

Common Core State Standards Grade Level Content (Middle and High School)
Interpreting Categorical and Quantitative Data
- Summarize, represent, and interpret data on two categorical and quantitative variables
- Interpret linear models

NCTM Principles and Standards for School Mathematics
Data Analysis and Probability Standards for Grades 9-12
Select and use appropriate statistical methods to analyze data
- for bivariate measurement data, be able to display a scatterplot, describe its shape, and determine regression coefficients, regression equations, and correlation coefficients using technological tools;

- identify trends in bivariate data and find functions that model the data or transform the data so that they can be modeled

Prerequisites
Prior to completing this activity, students should have studied correlations and regression analysis, including data transformations to achieve linearity.

Learning Targets
After completing the activity, students will be able to use real data to create a regression model, examine the model for linearity, and transform the data, if necessary, to create an appropriate model.

Time Required
The time required for this activity is approximately 60 minutes.

Materials Required
The students will need some device to conduct the regression, such as a calculator or computer. They will also need the classified ads from a newspaper or internet access. Note: Picking cars that are very fashionable, such as the Mini Cooper may result in a less rich lesson because the value of those cars is maintained even as they get older. The ideal car model to pick would be Honda, Toyota, and GMC. If the sample data does not require a transformation, then transformation questions should not be addressed. Perhaps students can be encouraged to try another car at that point. It should be noted that even if a transformation does not take place, a rich conversation about transformations can be carried out.

Instructional Lesson Plan

The GAISE Statistical Problem-Solving Procedure

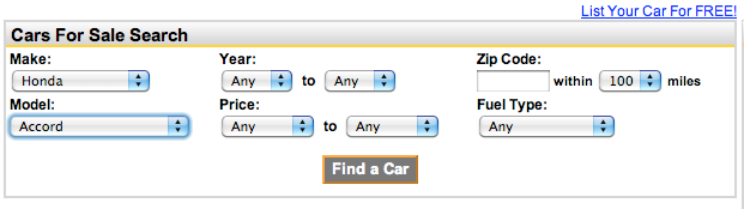
I. Formulate Question(s)

Begin the activity by passing out the "Used Car Regression Activity Sheet" and explaining that the goal of the lesson is for students to demonstrate how to create an appropriate regression model using real data. They will need to test the appropriateness of the model and transform the data to create a more appropriate model. In addition, they must interpret R^2, the intercept, and slope within the context of the problem.

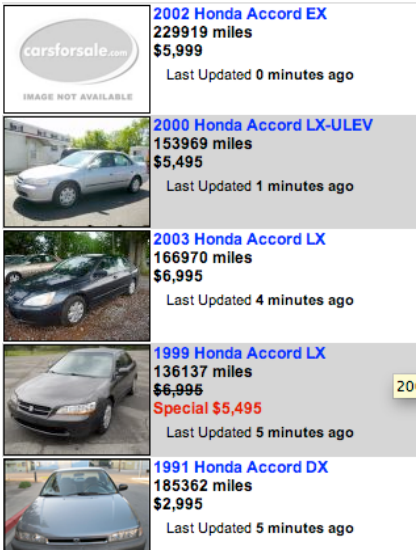
II. Design and Implement a Plan to Collect the Data

Use the classified ads from a local newspaper (or some online source) to collect the data on the following two variables – the age in years and the price of a specific car make and model (e.g., Ford Mustang or Honda Accord). The website http://www.carsforsale.com/ will allow the student to select the make and model, which will result in list of cars in order of most recent entry to oldest entry.

Here is an example of a search for a Honda Accord using http://www.carsforsale.com/.



After selecting the make and model, press "Find a Car." The following is a portion of the results obtained from a typical search. The year should be recoded by subtracting the year the car was made from the current year. For example, if this is 2014, 2012 would be coded as "2", 2011 as "3", etc.



III. Analyze the Data

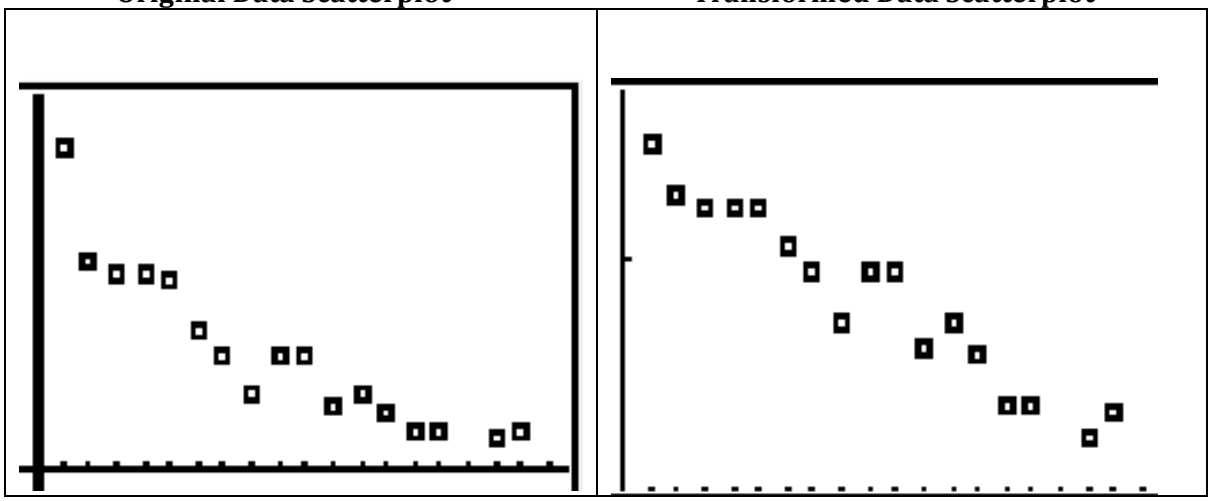
After inputting the data, complete the following tasks:

1. Search through the list and find 15 to 20 cars of the same make and model.
2. Using your selected device, enter the "Age of Car" in the first variable and the "Price" as the second variable. "Age of Car" should be recorded as current year minus model year. For example, a 2010 car is entered as 3.
3. Create a scatterplot showing the relationship between the two variables.
4. Run the regression and record the equation of the line and R^2.
5. Graph the residuals and determine whether the data need to be transformed. If the residuals look patterned, such as logarithmic, exponential, or parabolic, the data may need to be transformed.
6. Determine the appropriate transformation and transform the data. Experiment with various transformations.
7. Create a scatterplot for the transformed data.
8. Rerun the regression analysis and the residual plot for the transformed data.
9. Interpret the slope, intercept, and R^2 within the context of the scenario.

Sample Summary Report

Sketch the graph of the original and the transformed data.

Original Data Scatterplot **Transformed Data Scatterplot**

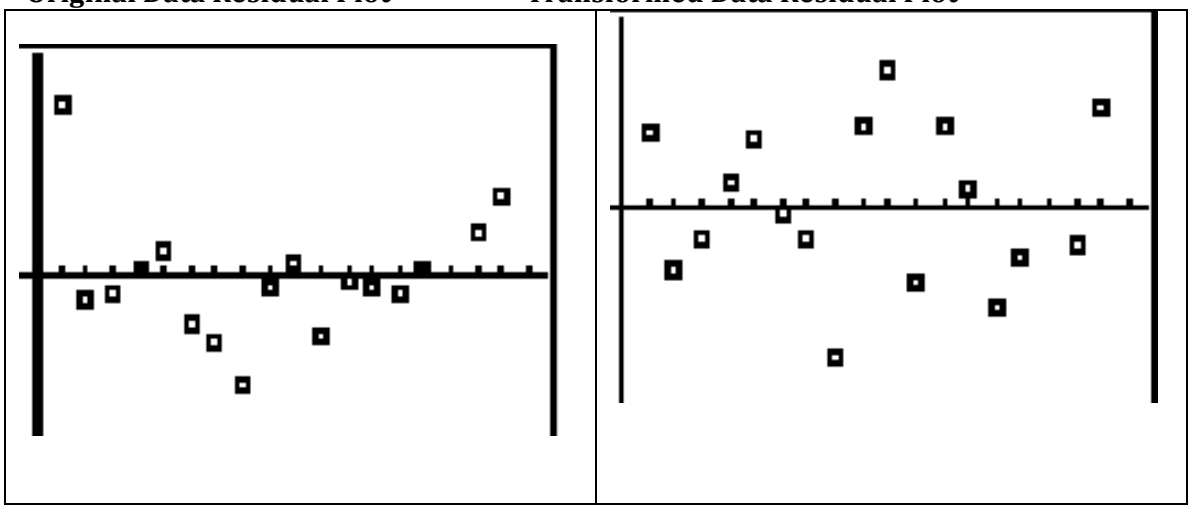


Complete the following table for the data on the original and transformed data. Interpret R^2, the slope, and the intercept in the context of these data.

Data Value	Original Data	Transformed Data
Model	Price = 19650.37533 – 1131.770 (Age)	Log (Price) = 4.401 - .058626 (Age)
R^2	R^2 = .834 Approximately 83% of the variation in price can be explained by the variation in year.	R^2 = .93576 Approximately 93.5% of the variation in price can be explained by the variation in year.
Slope	-1131.77 which means that the car will decrease in price by $1,131.77 with each increase of one year in age.	Each increase in one-year increase in the age of the car is associated with a -.058626 decrease in the log value of the price.
Intercept	19650.375 which means that a car that is 0 years old costs $19,650.77.	10^(4.40) = 25,118.86432 which means that a car that is 0 years old costs $25,118.86

Sketch the residual plots for the original and transformed data.

Original Data Residual Plot **Transformed Data Residual Plot**



Write a brief summary explaining the results of this analysis. Which model was better – the original or the transformed? What evidence do you have that supports your choice of best model? Explain R^2, the slope, and the intercept within the context of the problem.

The original scatterplot indicates that the relationship between the age of the car and the price of the car is not linear. The graph begins to curve at the base. In addition, the model only accounts for approximately 83% of the variation in price. The most compelling indication that the model may not be appropriate comes from the residual plot, which shows a distinct curved pattern in the residuals.

The data were transformed using a log transformation of the Y variable (Price). This transformation produces a much straighter plot and a better residual plot. In addition, the transformed model accounts for almost 94% of the variation, a substantial increase over the original model.

Write a brief summary statement about the relationship between the correlation, r, and the residual plots.

If the correlation is close to -1 or 1, the residuals are small.

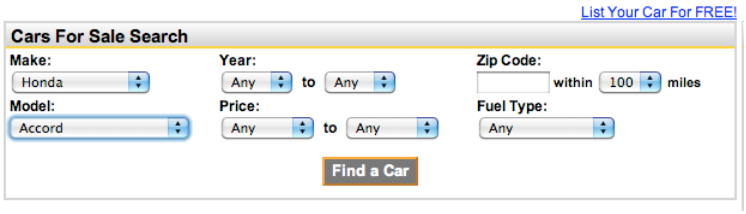
Write a brief summary statement about the relationship between the coefficient of determination, r^2, and the residual plots.

If r^2 is high, the residuals are small.

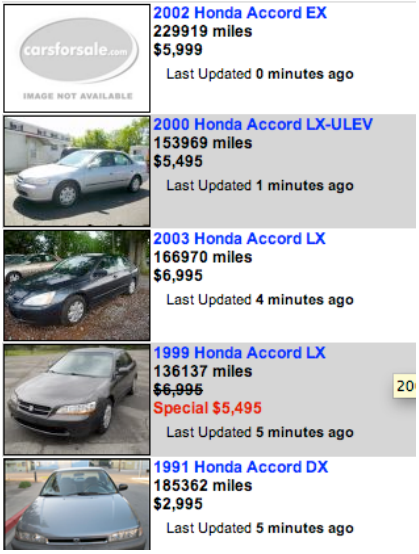
Used Car Regression Student Sheet

In this activity, you are going to estimate a appropriate linear model to predict the price of a used car from the age of the car. We will investigate the question: what is the relationship between a used car's age and price? Use the classified ads from a local newspaper (or some online source) to collect the data on the following two variables – the age in years and the price of a specific car make and model (e.g., Ford Mustang or Honda Accord). The website http://www.carsforsale.com/ will allow the student to select the make and model, which will result in list of cars in order of most recent entry to oldest entry.

Here is an example of a search for a Honda Accord.



After selecting the make and model, press "Find a Car." Following is a portion of the results obtained from a typical search.



Analyze the Data

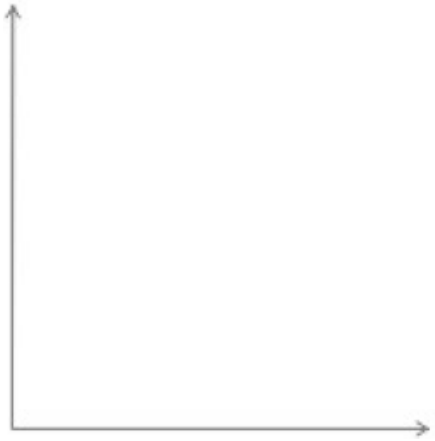
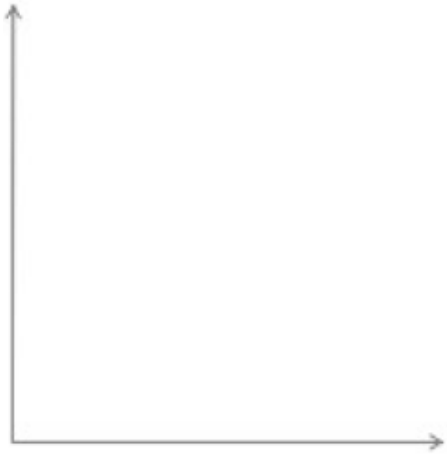
After inputting the data, complete the following tasks:

1. Search through the list and find 15 to 20 cars of the same make and model.
2. Enter the "Age of Car" in the first variable and the "Price" as the second variable. Age of Car should be recorded as current year minus model year. For example, a 2010 car is entered as 3.
3. Create a scatterplot showing the relationship between the two variables.
4. Run the regression and record the equation of the line and R^2.
5. Graph the residuals and determine whether the data need to be transformed.
6. Determine the appropriate transformation and transform the data.
7. Create a scatterplot for the transformed data.
8. Rerun the regression analysis and the residual plot for the transformed data.
9. Interpret the slope, intercept, and R^2 within the context of the scenario.

IV. Interpret the Results

Summary Report

Sketch the graph of the original and transformed.

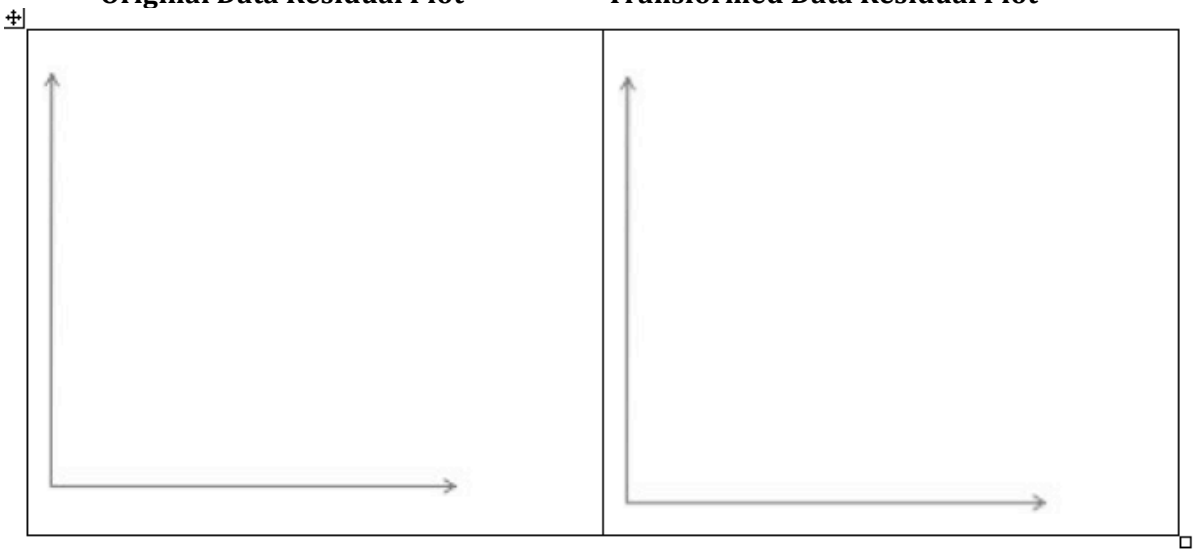
Original Data Scatterplot	**Transformed Data Scatterplot**
	

Complete the following table for the data on the original and transformed data. Interpret R^2, the slope, and the intercept in the context of these data.

Data Value	Original Data	Transformed Data
Model		
R^2		
Slope		
Intercept		

Sketch the residual plots for the original and transformed data (if applicable).

Original Data Residual Plot **Transformed Data Residual Plot**



Write a brief summary about the relationship between the two variables. Which model was more appropriate – the original or the transformed? What evidence do you have to support your choice of best model?

Write a brief summary statement about the relationship between the correlation, r, and the residual plots.

Write a brief summary statement about the relationship between the coefficient of determination, r^2, and the residual plots.