STAT 140: Week Three Catch Up

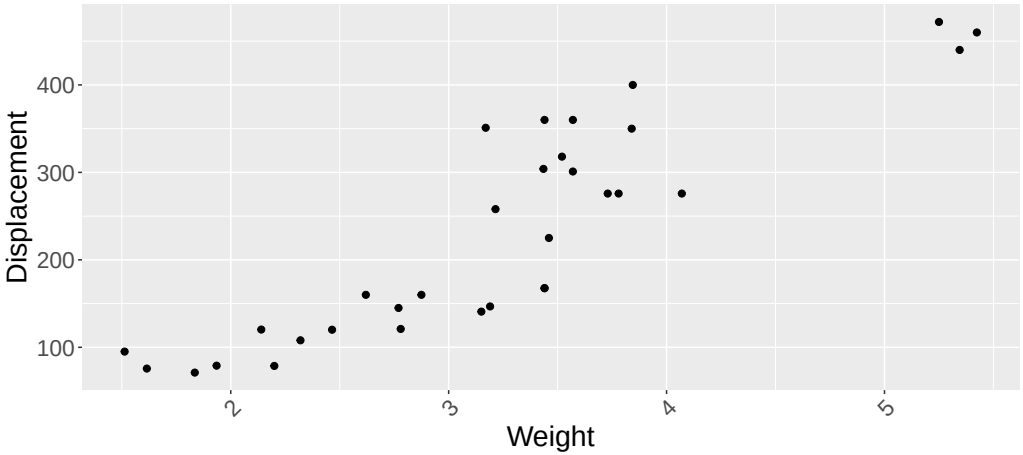
09/15/2021

My hope is that after Week 3 (09/13-09/17), you will be able to:

My hope is that after this week, you will be able to:

- Summarize relationships between variables, in graphs and with statistics.
- Interpret statistics for relationships between variables.
- Fit a least squares regression line.
- Determine the appropriateness of a least squares regression line.

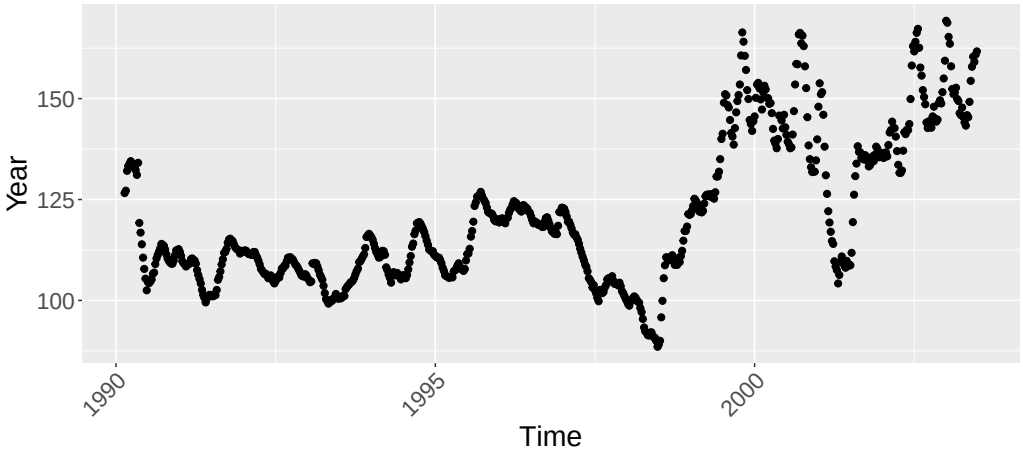
Lab Exercises: `mtcars` 1



Lab Exercises: `mtcars` 2

- What I said: Strong positive linear relationship with a few unusual values (but you can make a case for no unusual values)
- What is true: Strong positive non-linear (quadratic) relationship!!

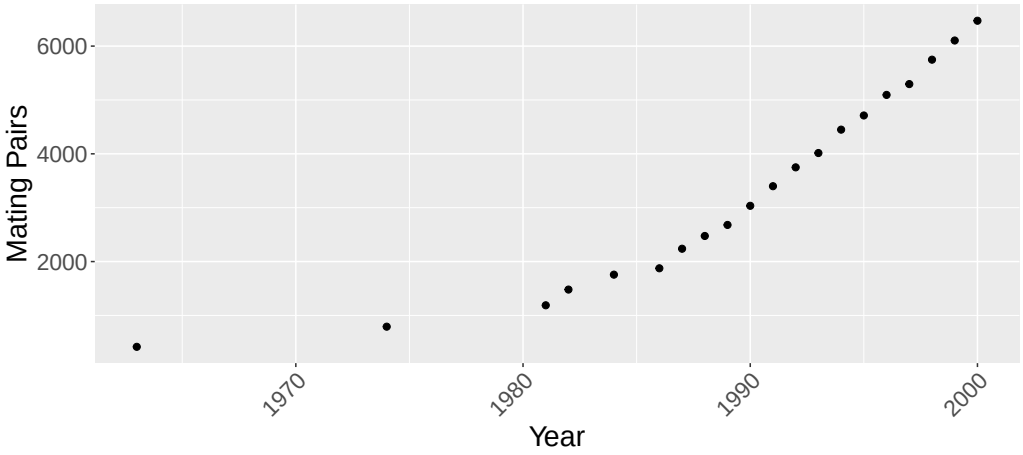
Lab Exercises: gasprice 1



Lab Exercises: gasprice 2

- What I said: Strong positive nonlinear relationship with no unusual values
- What is true: Time Series (uses a whole different set of statistical tools)

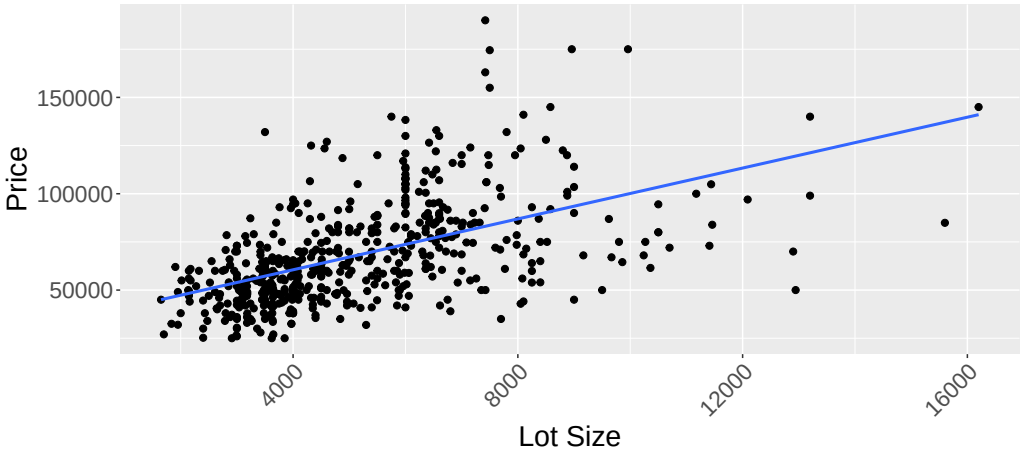
Lab Exercises: `eagles` 1



Lab Exercises: `eagles` 2

- What I said: Strong positive linear relationship with no unusual values (but you can make a case for some unusual values)
- What is true: Strong positive non-linear (exponential) relationship with no unusual values!!

Lab Exercises: `HousePrices` 1



Lab Exercises: `HousePrices` 2

- Moderate positive linear relationship with no unusual values (but you can make a case for some unusual values)

Notation

Question: could you go over the difference between population and sample, mathematically, so we see why the notations are different?

Characteristic	Parameter	Statistic
Mean	—	$\bar{x}$
Standard Deviation	—	s
Proportion	—	$\hat{p}$
Correlation	—	R
Intercept	β_0	b_0
Slope	β_1	b_1

Form 1

Question: Would an exponential graph be considered non-linear?

Yes, an exponential graph is one example of a nonlinear relationship! You can also consider quadratic graphs, time series, and more to be nonlinear.

Question: In some cases, I find it very difficult to know whether it is a linear relationship or an exponential relationship.

Yes–it can be really difficult to see. For now, I won't ask you to deal with any non-linear relationships, and remember that we will learn more tools for detecting linearity later on in the course.

Form 2

Question: Will we learn how to model non-linear relationships and choose the best fit from different types of relationships?

Possibly! It depends on how far we get this semester–we might reach a point where you can learn some tools to start dealing with non-linear relationships. But, there are quite a few non-linear forms, so we definitely won't learn them all! You would also learn this in STAT-242, if you continue on. As for the best fit, yes, we will definitely learn more about that!

Strength 1

Question: I know that strong correlations are close to -1 or 1, while weak ones are close to 0, but are there particular values associated with qualitative descriptions of correlation strength? For example, can a specific range of correlation values be called weak, while others can be called moderate or strong?

There's no official range, but here's usually what I say:

- Weak: $0 \leq |R| < 0.3$
- Moderate: $0.3 \leq |R| < 0.7$
- Strong: $0.7 \leq |R| < 1$
- Perfect: $|R| = 1$

Strength 2

Question: Is correlation a good way to measure the relationship of non-linear trend? The calculation of correlation seemed to be designed for linear relationships. But if not using correlation, is there an alternative way or quantity that can be used to measure non-linear trend?

I would say no–if your relationship isn't linear, don't use correlation to measure the strength. There are some ways you can measure the strength of some non-linear trends, but we probably won't learn about them in this class.

Unusual Values

Question: Determining whether a scatterplot has any unusual values seems a bit subjective at times. Are there any particular criteria that indicate an unusual value?

Yes! But we won't learn them until later–again, for now, just eyeballing is fine.

Linear Regression Equations

Question: Why does the intercept come first in the linear equation we use?

Sort of by convention–this is just how most statisticians do it! It doesn't really matter and I would accept either order on an assignment.

Question: Does it make sense to analyze its linear regression formula in a non-linear dataset?

This is a great question! I think no. First of all, if you get even a little bit further on in statistics, you will learn some tools that can help fix non-linear relationships relatively quickly, and if you can fix it, you should. Second of all, we will learn some methods for drawing stronger conclusions about linear relationships later in the semester that rely on assumptions, especially the assumption of linearity–if you make an invalid assumption, you can't really trust the results! That would be a lot of work for something you can't have confidence in.

Residuals

Question: Could you call the residuals the error? Or is it just one type of error?

Residuals represent a general error term that covers any type of error–in this class, we will only focus on general error, but there are a lot of statisticians researching different types of error (me!).

Question: Do you calculate the residual of individual points on the graph or is there a way to calculate or estimate a residual margin?
I'm not quite sure what you mean by residual margin, but we can calculate the residuals for each individual point. Typically, we aren't necessarily interested in these values by themselves–statisticians usually consider them all together. We will learn more about this later on in the course.

R

Question: I want to know the meaning of `vjust` and `hjust` in the following code:

```
theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1))
```

`hjust` and `vjust` stand for horizontal and vertical justification–sometimes you change these arguments to make the axis text easier to read.

Question: I am a little bit confused on how to assign names to variables and then use them in code. This is especially confusing in more complex things, like writing the equation for a line. Will we be asked to do this? Could we go over finding the equation of a line in R?

You should definitely know how to save the output of the `lm()` in an object, but as for the slope/intercept like we did in lab–I probably would not ask you to do that.

Journal Questions and More from Week 2

Question: One of the articles for the journal options displayed a stacked bar graph. When is it appropriate to use those?

Stacked bar charts show the distributions of a categorical variable for two or more groups.

Question: How can we identify the relationship between correlation and skew based on a histogram?

There's no relationship between correlation and skew! Correlation is used to talk about a relationship between two variables, and skewness is used to talk about the distribution of one by itself.

Test 1

- You will take your first test sometime next week, from Thursday 9/23-Sunday 9/26.
- The test will be taken in Moodle, on your own time! It should not take more than an hour, but you can take as long as you need.
- However, the test is closed note/closed book–you should not be looking at any outside resources while your test is open.
- The test will cover content from class, including lab, and homeworks and progress checks. It will be all short answer type questions.
  - You will not have to review any journal content.
  - You will not have to decipher a scatterplot, or any others.
- The best way to practice is to review old progress checks (which I have reopened), and I will offer more practice questions next week.