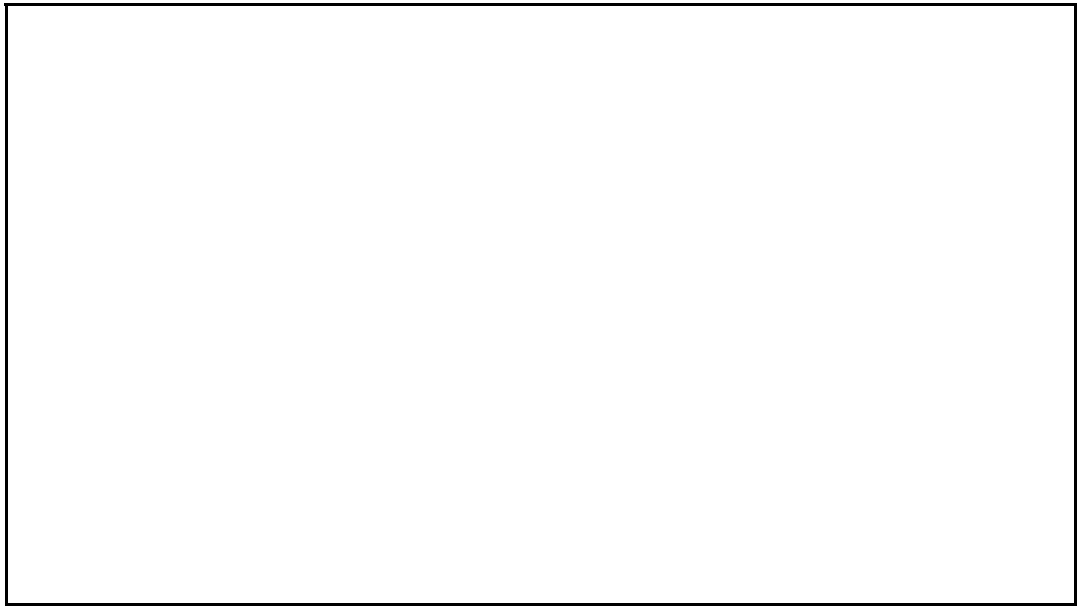
THE LINE OF BEST FIT



0:00 / 9:10 1.0x

SPEAKER: MICHAEL J. MAHOMETA, Ph.D.

We've learned that we can actually fit a line - a linear model or function -

to some data.

But how does this actually happen?

How do we decide on a model?

How does software decide on a model?

Were there linear models before software was around?

How does this line - the line that best fits

our data, the line that is our model - come about?

Well, it all has to do with something called residuals.

First let's take a look at our question.

I was out riding my bike with my family recently,

and I happened to notice other children on their bikes at the same time.

And some of them were wearing helmets and others were not.

Having just bought a helmet, I knew that the cost could be pretty high

and wanted to further investigate.

Let's start with a scatterplot.

Here's some simulated data on 15 elementary schools:

the percent of children receiving free lunches

and the precent of children wearing bicycle helmets for each school.

We're using the free lunches idea as a surrogate measure

of socioeconomic status.

The scatterplot looks good - no outliers, no clustering,

no non-linear shape to it.

Given this, the linear model seems like a good choice to use.

It makes sense that we could draw a straight line through the data.

But where should we start?

Where should that line fall?

Help

Remember that, when we were looking at the correlation,

there was something great about the scatterplot.

That we could use the mean of x and the mean of y

to help out in the calculation of the Pearson Correlation Coefficient.

Well, it turns out - as long as we're talking about a linear model

- we can use those values of the mean of x and the mean of y again.

It turns out that any linear function line

we draw though correlated data MUST pass through the value of the mean of x

and the mean of y.

In our case that point is (36, 28.5).

Knowing this, we can draw any line we want,

but unfortunately there's an infinite number of lines that are possible.

How do we know which line is the best line?

Well, we look at residuals.

Let's draw a line here - starting at an intercept value of 60

and going through the point of the mean of x and the mean of y.

Notice for every point on the graph, our line is wrong by some distance

- and that distance is different for every point.

For example, the distance that our line is wrong for this school

is this particular amount, while for this other school,

our line is wrong by only this amount.

This amount wrong is called a residual, and can formally

be defined as the value of y minus the predicted value of y.

Let's see what all the residuals for every point are.

Notice, some are above the line and some are below the line.

So some will have a positive residual and some will have a negative residual.

If we add up all the residuals, just like if we

were to sum all the deviations of a distribution,

we would actually get zero.

So we square each residual and then sum.

And we get our sums of squared residuals value.

This gives us a measure of appropriateness

of the line we just drew - how well it works.

Help

Unfortunately, the only thing we can compare that value to

is another sums of squared residuals value from another line.

So, let's take a look at a simulation - we'll fit a bunch of lines

and calculate the sums of squared residuals for each.

Notice at that there's a point on our sums of squared residuals

graph that bottoms out.

This is where the residuals from the line - the error - is the lowest.

For our data, that lowest value occurs at a linear function of y=47+(-0.51x)

Now here's the cool thing.

If we solve for the line of best fit using the calculation method,

we get a linear function that's very close to what we just

found through simulation.

What does this all mean?

It means that the line of best fit for a linear

function

- the line that we get from software - is

the line that has the lowest value of sums

of squared residuals of all the possible lines we could draw.

Help

So to answer our question, the linear model that best fits the data

of socioeconomic status and bicycle helmet wearing among elementary age

children is f(x) = 46.91+(-0.51x).

Speaking of sums of Squared residuals - there's

this thing called "coefficient of determination"

that's associated with the Pearson Correlation Coefficient.

It's simply the correlation coefficient squared.

For the current relationship between socioeconomic status and bicycle helmet

wearing behavior, that value is 0.713.

Let's investigate where this value actually comes from.

First let's start with a little thought exercise.

Imagine you wanted to predict the percent of children wearing bicycle

helmets, but you didn't know the percent of children taking free lunches.

Could you predict the the percent of children wearing bicycle helmets?

Sure - do you know how?

You would use the mean of the percent of children

wearing bicycle helmets for the current sample.

Help

And you would obviously be wrong - but how wrong?

Well, we could find out a "base level of incorrectness"

by using the idea of residuals again.

Here's the scatterplot again, and here's the mean value

of the percent of children wearing bicycle helmets.

And here are the residuals of that line (y - y bar).

We could square then sum these residual values and get some value - it's 3639.

This value again doesn't mean much on its own, but let's just keep it handy.

Now, in fact, we know the percent of children taking free lunches,

and we know that there's a relationship to percent

of children wearing bicycle helmets.

And we just fit a model to that relationship.

We also know that that model is not perfect.

There's some error to the model's prediction value

- and we know this as the residual.

Remember the sums of squared residual for the line of best fit

we recently found?

It was 1043.

Let me ask you this: How much "error" in the prediction of the percent

of children wearing bicycle helmets, as measured by the sums of squared

residual did you decrease or SAVE by using the model over the sample mean

value?

A little quick math gives us 2596.

Let's put this number in the context of the original "base error"

by using a ratio.

Now go ahead and solve for this ratio, I'll wait.

Did you find it?

Did you find the ratio?

Guess what, it matches the value for the "coefficient of determination."

Now, itnterestingly, the "coefficient of determination"

is also called the "proportion of variance accounted

for" because that's what it IS.

It's the amount of error that we can account for in the variable of interest

- the dependent variable - by knowing the model of the relationship

with it and our independent variable.

So to continue to "fill out" our original answer,

our model of f(x) = 46.91+(-0.51x), has a "proportion of variance accounted

for" of 0.713 or, 71.3% of the variance of the error in the prediction

of the percent of children wearing bicycle helmets,

can be accounted for by knowing the percent of free lunches for every

school.

Now we'll eventually see how we can use this value -

the "proportion of variance accounted for"

- to help in determining which model (the linear, the exponential,

or the logistic growth model) works better with our specific data.

Help

Comprehension Check

1. What is a residual? Select all that apply.

(1/1 point)

- [ ] the average distance between any data point and the regression line
- [ ] the difference you get when you subtract a data point from the next closest data point
- [x] the distance between a data point in a scatterplot and the line of best fit
- [x] $e=y-\hat{y}$

Check Show Answer

Help

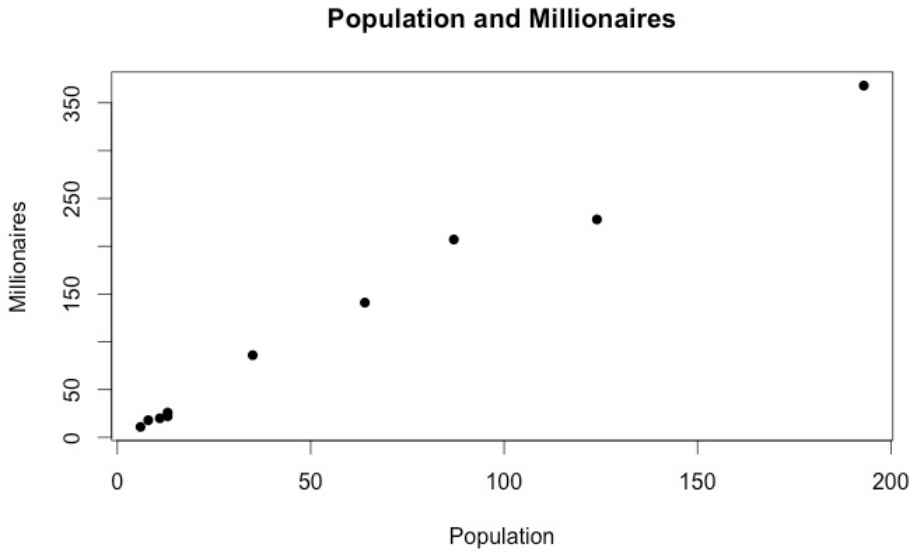
2. Do states with higher populations have more millionaires? Here is data from 2008. The variable labeled "Population" in the table and scatterplot will be referred to as "State.Population" in the questions that follow in order to avoid confusion with the meaning of "population" as a concept in statistics.

State	Millionaires *(in thousands)*	Population *(in hundreds of thousands)*
Connecticut	86	35
Delaware	18	8
Maine	22	13
Massachusetts	141	64
New Hampshire	26	13
New jersey	207	87
New York	368	193
Pennsylvania	228	124
Rhode Island	20	11
Vermont	11	6

Using linFit(), the following linear model is found:

Millionaires = 6.296 + (1.921 * State.Population)

Help



(2/2 points)

2a. What is the correlation between Millionaires and State.Population? *(Round to 3 decimal places.)*

- -0.454
- 0.763
- 0.992 ✔
- 1.921

2b. What is the coefficient of determination? *(Round to 3 decimal places.)*

- 0.015
- 0.763
- 0.985 ✔
- 0.992

Help

Check Hide Answer

Help

About edX

About

News

Contact

FAQ

edX Blog

Donate to edX

Jobs at edX

Follow Us

Twitter

Facebook

Meetup

LinkedIn

Google+