AGAINST ALL ODDS
EPISODE 11 – "FITTING LINES TO DATA"
TRANSCRIPT

FUNDER CREDITS

Funding for this program is provided by Annenberg Learner.

INTRO

Pardis Sabeti
Hi, I'm Pardis Sabeti and this is *Against All Odds*, where we make statistics count.

Scatterplots are a great way to visualize the relationship between two variables… such as ocean temperatures and coral reef growth. As temps go up, new growth goes down. Or between femur bone length and overall height… as seen in this clip from *Bones*, Fox's show about a forensics team.

> **Temperance Brennan**
> A lot of jewelry. Male. Thighbone suggests he was tall. I.D. bracelet….

Pardis Sabeti
Forensic anthropologists know from years of collecting data that there's a relationship between the length of your thighbone and your total height.

> **Temperance Brennan**
> …that makes a difference?
>
> **Seeley Booth**
> Facts of life, Bones…

Pardis Sabeti
In this case the femur length is what influences the total height. Remember that the explanatory variable – femur bone length – goes on the *x*-axis – and the response variable – total height – goes on the *y*-axis.

The data points appear to fall pretty much along a straight line… longer thigh bones and taller heights look like they go together in a linear relationship. Statisticians call the line that describes how the response variable changes with the explanatory variable a regression line.

Any straight line can be described by the equation $y = a + bx$. a is the y-intercept – the spot where the line hits the y-axis, where x equals 0. b is the slope of the line – how much y changes when x goes up by one. x and y are the data points you're plotting – in this case, the bone and height measurements.

Sizing up the data points, we can just eyeball where the line should fall… but there's a statistical technique to figure out how best to fit a line to the data. Once we have that line, we can use it to make predictions. That's how Brennan on *Bones* can estimate an unidentified crime victim's height even when she has an incomplete or damaged skeleton – just using the thighbone measurement.

> **Zack Addy**

Approximately 6 foot 7. Right-handed.

Seeley Booth
Six foot 7?

Temperance Brennan
This man was dead for several hours before the train hit him.

Pardis Sabeti
The Colorado Climate Center uses regression lines in a less sinister scenario – to forecast the state's seasonal water supply. Farmers, city planners, businesses; it's important to all of them to know how much water is going to be available each year so they can plan accordingly.

Nolan Doesken
The sooner you know that, the better you can strategize for planting crops, for managing landscapes, golf courses, for knowing what water can be stored in reservoirs, for understanding what water will be available to meet the interstate compacts on the various rivers.
The big question… is…"how can we predict the water supply we're going to have as far ahead of time as possible?"

Pardis Sabeti
Most of Colorado's water comes from melting mountain snows in the spring. So climatologists have developed a model based on two types of data: the amount of winter snowpack in the high elevations and the resulting volume of water that flows out of the mountains throughout the summer.

Nolan Doesken
What you really like is to have many years of that in the past so that you can see how the snow water content in the mountains over time has related to the amount of water flowing down the river during the late spring and summer.

Pardis Sabeti
So Colorado's Natural Resources Conservation Service heads into the Rockies to collect that data.

Greg O'Neill
They've got hundreds of sites in the Colorado mountains and they monitor those on a monthly basis. They do that by identifying sites that they return to, year after year after year. And they collect a minimum of ten depth and quantity subsections on the snowpack and then they average those to come up with the total for that particular area.

Pardis Sabeti
So that's one part of the model: the water that's stored as snow all winter long. The second set of data the climatologists need is how much water actually ends up running downstream.

> **Greg O'Neill**
> The U.S. Geological Survey operates a network of about 300 gauging stations in Colorado.
> At each one of those locations we have a recording device that continuously logs the height of the river and then we come up with a continuous record of discharge at each one of these 300 sites.

Pardis Sabeti
Let's graph these data points on a scatterplot. You can see that in the years when our explanatory variable – the snowpack – was high, the spring runoff – our response variable – was too. It looks like a pretty strong positive linear relationship.

Of course in the real world, all data points don't fall on an exact, precise line. So we need a technique to figure out the regression line that minimizes the vertical distances of our data points from the line.

Let's zoom in on just these three points from our Colorado water data. We want to find the best-fitting line, with the most points as close as possible to it. This vertical distance of a point from a line is called a residual. As we shift the line, some residuals get smaller while others get larger. They can't all be small at the same time, so we need to find that sweet spot where we've gotten them as a group as small as we can.

Statisticians use a method called least-squares to do that. Since some of the residuals are positive – above the line – and some are negative – below the line – we square them, which makes them all positive. If you add up the squared residuals, the bigger the sum, the more the line misses the points. So we want to make that sum as small as possible. Software can do the math for you and come up with the equation "y-hat" = $a + bx$ to describe our line.

We say "y-hat" to emphasize that we're talking about the predicted value of y, not a measured value from our data set. Now we have our regression line and can use it to make predictions! Remember, the plotted points are our actual data, and the regression line predicts the y-value for any x.

Back to the Colorado water data….

Statistical software gives us the equation for the regression line. The slope of 1,941 predicts that for every one inch increase in snowpack, the runoff increases

by 1,941 acre feet. The *y*-intercept is negative 7,920 … which would seem to say that if the snowpack was zero, the runoff would be around negative 8 thousand acre feet. Obviously that doesn't make sense, and it's a good reminder that you can't extrapolate from the regression line too far outside the range of the observed data. Keeping that limitation in mind, though, the regression line can be very useful for Colorado water users.

> **Nolan Doesken**
> We have found that it really does predict the amount of stream flow that we're going to have months in advance.

Pardis Sabeti
If you know that this winter, the Rockies saw 30 inches of snowpack, you can look at the line to predict how much water is going to flow into the system in the spring. Here's the spot on the line; and plugging in the numbers to the line's equation gives us a forecast of 50,310 acre feet of melt water. That's a number that can help water resource managers plan for the year ahead.

The regression line works well to predict the Colorado water supply because the relationship between snowpack and river flow is linear. If the relationship you're trying to describe actually has a curved pattern, a straight regression line won't be a good fit to the data. For instance, check out this scatterplot showing the curved relationship between an alligator's length and weight. Trying to use a regression line in a case like this won't make the most accurate predictions.

One way to assess how well a regression line fits the data is to make a residual plot. That's a scatterplot of all the residuals against the explanatory variable. It's as if you turned the regression line horizontal. If the dots in the residual plot appear randomly scattered with no strong pattern—like they do in the Colorado water case—both above and below the line, then the regression line has nicely captured the pattern in the data and a linear model is a good choice to describe it. But if the residual plot looks at all curved – like in the case of our alligators – then the data pattern is curved and the linear model is the wrong one to use.

Probably best to steer clear of those alligators… especially if they're hungry. You wouldn't want to become a meal that would add to their weight, no matter what the relationship to their length!

For *Against All Odds*, I'm Pardis Sabeti. See you next time!

PRODUCTION CREDITS

Host – Dr. Pardis Sabeti

Writer/Producer/Director – Maggie Villiger

Associate Producer – Katharine Duffy

Editors – Brian Truglio
-- Jared Morris
-- Seth Bender

Director of Photography – Dan Lyons

Sound Mix – Richard Bock

Animation – Jason Tierney

Title Animation – Jeremy Angier

Web + Interactive Developer – Matt Denault / Azility, Inc.

Website Designer – Dana Busch

Production Assistant – Kristopher Cain

Teleprompter – Sue Willard-Kiess

Hair/Makeup – Emily Damron

Additional Footage:

- FOX Broadcasting Company
- National Parks Service
- Pond5/yofitofu
- Pond5/Abaget
- City & County of Denver – Media Services
- iStock/alptraum
- Coast Learning Systems

Music
DeWolfe Music Library

Based on the original Annenberg/CPB series *Against All Odds*,
Executive Producer Joe Blatt

Annenberg Learner Program Officer – Michele McLeod

Project Manager – Dr. Sol Garfunkel

Chief Content Advisor – Dr. Marsha Davis

Executive Producer – Graham Chedd

FUNDER CREDITS

Funding for this program is provided by Annenberg Learner.

For information about this, and other Annenberg Learner programs, call 1-800-LEARNER, and visit us at www.learner.org.