Actua's AI Activities Series

Activity 3

Regression Analysis: Making Predictions Using Data

Actua | 111 Murray Street, Ottawa, ON K1N 5M5 | www.actua.ca | 613 234 4137

Youth · STEM · Innovation
Jeunesse · STIM · Innovation

Regression Analysis: Making Predictions Using Data

If you're accessing this activity directly, did you know there are eight other activities in this series up on our website? If you find yourself unfamiliar with any of the AI concepts and terminology introduced in these activities, please refer to our AI Glossary. These activities also follow a space exploration narrative when done in order. It is recommended to complete the activities in order but they can also be done on their own.

You and your group mates are astronauts and scientists aboard the Actua Orbital Station. Unfortunately, your station just got bombarded by magnetic rays and your electronics have begun to shut down! The only one who can save you is the station's AI, DANN. DANN stands for Dedicated Actua Neural Network, and it's gone a little loopy. Brush up on your technical skills, learn about AI, and save yourself and your crewmates!

The space objects classifier is working again, thanks to your efforts in "Decision Trees: Classifying Space Objects". Now, DANN is currently busy running a diagnostic to determine what needs to be fixed, and we can't help until it finishes. Meanwhile, you've discovered some data from one of the old experiments aboard the Station. The experiment was studying the effect of microgravity on astronauts. So, Mission Control has encouraged you to take a look at this experiment's data and see if you can do a preliminary analysis. Once that's done, Dann will be finished its diagnostic, and we can start fixing it in "Sort Things Out: Exploring Image Classification"!

Activity Summary

In this activity, participants will analyze data to determine the relationship between two variables for the purpose of making predictions. Participants will make estimates by visually analysing data. Then, they will use a technique called regression to create a model that can make predictions from the data. Finally, participants will reflect on the process of creating and evaluating prediction models and discuss how machine learning could be used to create similar models.

Developed by Actua, 2022.

Delivery Environment	Activity Duration	Intended Audience
Computer lab/Classroom with computers	60-75 minutes	Grades 9-12 (Ages 13-18)

Achievement Goals

Learning Goals

Learning goals are statements referring to the understanding, knowledge, skills or application participants acquire during the activity. **Following this activity, participants will:**

- **Analyze** the relationship between dependent and independent variables.
- **Understand** the importance of prediction models and their impacts.
- **Explore and analyze** the varying accuracy of statistics and mathematical models.

Logistics (Timing, Group Size, Materials)

Section Title	Time	Group Size	Materials
Opening Hook: Guessing the Future	5 minutes	*Entire Group*	
Activity 1: What is regression?	15 minutes	*Entire Group*	**Whole Group** • Data and scatter plot (or access to it on a computer) • Pen/pencil

Section Title	Time	Group Size	Materials
Activity 2: Creating a Model	20 minutes	*Individually or in small groups*	**Whole Group** • Computer with Internet access • Interactive sketch Line of Best Fit
Activity 3: Testing Predictions	20 minutes	*Individually or in small groups*	**Whole Group** • Computer with Internet access • Interactive sketch Line of Best Fit • Printed calculation table (or access to it on a computer) • Pen/pencil • Calculator
Reflection & Debrief	10 minutes	*Entire Group*	

Safety Considerations

Safety considerations have been provided below to support safety during this activity, however they are not necessarily comprehensive. It is important that you review the activity and your delivery environment to determine any additional safety considerations that you should be implementing for the delivery of these activities.

Online Safety

Some components of this activity require the use of devices connected to the internet.

- Facilitators should review the provided videos and read/explore provided websites and materials to determine if they are suitable for their participants.
- Where applicable, facilitators should remind participants to stay on task and only use links provided within this activity.

Activity Procedure

Opening Hook: Guessing the Future

Prediction is the act of making a guess about the future, based on present information. Machine learning can be used to train predictive models for a variety of tasks that can have a huge impact on our lives. As a group, consider the following questions:

1. Where could predictive models be used in our society? What could they predict?
    a. Examples could include weather forecasting, disease spread, volcanic eruptions, earthquakes, financial or commodity markets, animal migration, climate.
2. What makes a predictive model “good”? Why is it important for a predictive model to be good?
    a. A predictive model would be considered “good” if its predictions were accurate.

b. If a predictive model makes bad predictions, it would be unhelpful for decision making (or may even be dangerous if it causes the wrong decision to be made).

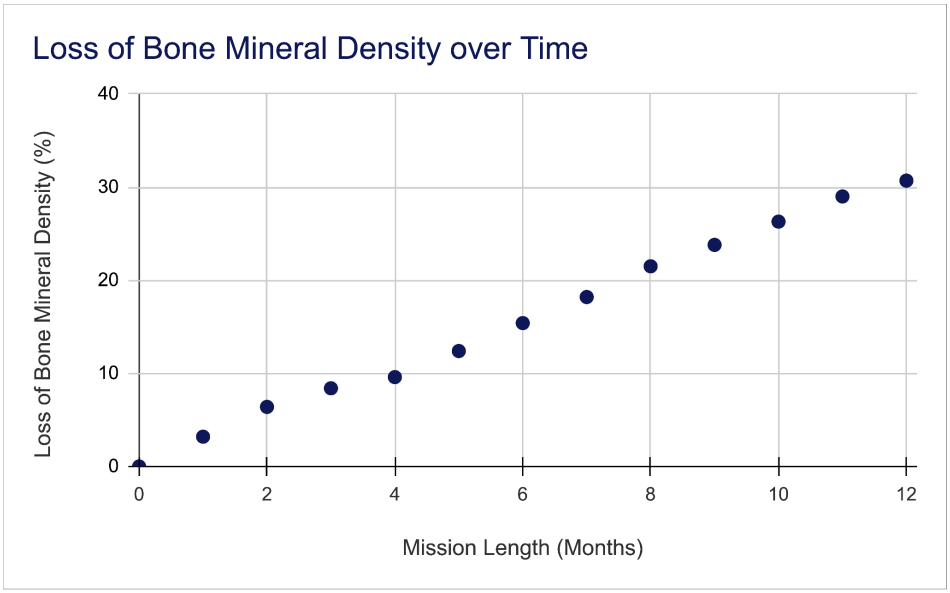
Section 1: What is Regression

To make good predictions, we need to collect data related to the predictions that we want to make. If we're interested in looking at the health impacts of long space missions, we need to collect data related to the health of our astronauts, and then analyse it to look for patterns. Below is a set of data generated from the medical examinations of the last group of astronauts on the Actua Orbital Station.

Mission months	Loss of bone mineral density (%)
0	0.0
1	3.2
2	6.4
3	8.4
4	9.6
5	12.4
6	15.4
7	18.2
8	21.5
9	23.8
10	26.3
11	29.0
12	30.7

This is an average of the data from each of the 6 astronauts that were part of that one year mission. The amount of loss is relative to their starting bone mineral density (so, it starts at 100% bone density for each of them in month 0). *How can we use this data to try and predict what would happen on a longer mission?*

First, let's see what this data looks like:



Based on that plot, we might hypothesize that there is a relationship between the amount of time in space (mission months) and the impact on our astronauts' health (as shown by the amount of bone mineral density loss). { In small groups / As a large group}, consider the following questions:

1. How could you describe the relationship between mission months and bone mineral density loss?
    a. The data in the scatter plot looks like it would fall approximately on a line, so we could call this kind of relationship linear.
2. How could you use this data and the plot above to predict the impact on astronaut health beyond 12 months?
    a. You could draw a line through your data points and then project the line past 12 months
3. How much bone mineral density loss do you think would have occurred by the end of an 18-month mission? What about a 24-month mission?
    a. If the linear relationship holds, you can predict the amount of loss by looking at the difference between month 6 and month 12 (approximately 15) and add that on to the month 12 loss to get 18 months (30 + 15 = 45), then add it again to get 24 months (45 + 15 = 60).

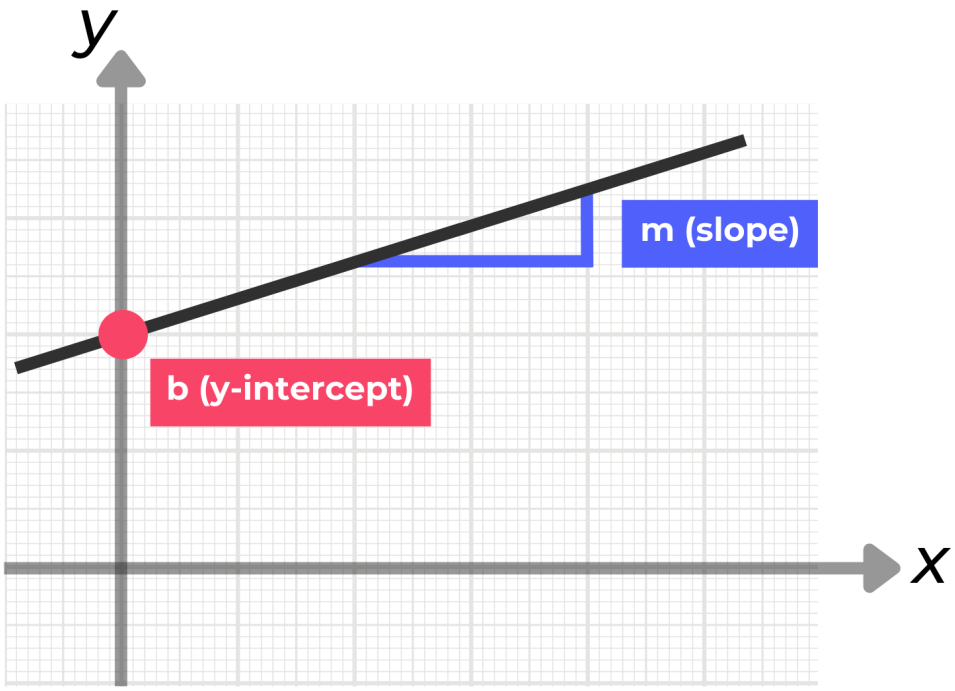
b. For 24 months, you could also take the difference between 0 and 12 months (30) and add it onto itself (30 + 30 = 60).

Section 2: Creating a Model

With our scatter plot data, we can make estimations by quickly analyzing the data visually, but what if we wanted to express this relationship mathematically? We can use a technique called “regression” to come up with a model that expresses the relationship between time in space and how it impacts the bone mineral density of our astronauts. Since we’ll be expressing this relationship in the form of a line, we’re specifically using “linear regression”, though other types exist for modelling more complex data. Linear regression works like this:

1. Find a line through your data that tries to get as close as possible to each data point.
2. Calculate how “good” of a fit your line is by comparing how far away each data point is from the line.
3. Make an adjustment to your line and repeat step 2 to see if it improves or worsens the fit of your line to the data.
4. Keep repeating step 3, making adjustments until you can’t improve the fit of your line to your data.

Remember that the equation of a line can be expressed as **y = mx + b:**



$$y = mx + b = 0.5x + 2$$

where **m** is the slope of the line and **b** is the height of the line when it crosses the y-axis (the "y-intercept" i.e. the value of y when x = 0). For a regression line, there's a small change from **y** to **ŷ** to indicate that this is a predicted value, and not a recorded one. For our model, we could also express this as:

(ŷ: Predicted bone mineral density loss) = (Slope: Rate of loss) × (x: Months in space) + b

The two underlined terms in that expression, slope and b, are the values that we can change to tune the fit of our regression line. **Slope** affects the "angle" of our line and **b** slides our line up and down the chart. "Months in space" is our independent variable and by picking a number for this value, we can use the equation to get a value for "predicted bone mineral density loss", our dependent variable.

With this in mind, we're ready to pick starting values for our slope and b. These can be anything that we want, but it's probably a good idea to use our data as a starting point and work from there. For example, if you drew a line through the first (0, 0) and last datapoint (12, 30.7), you could get the slope by calculating:

(y2 - y1) / (x2 - x1) = (30.7 - 0) / (12 - 0) = 2.56

Likewise, if we're using the first data point at (0, 0), then we'd set b = 0. So our first model could be:

(ŷ: Predicted bone mineral density loss) = 2.56 × (x: Months in space) + 0

But you can start with any two data points to get your own values of slope and b. In the interactive sketch below, you can place two points by clicking on two locations in the grid. This will tell you the slope and b for the line that is drawn. You'll also notice some yellow lines which represent the "residual" for each data point. Residuals will be discussed in the next part.

Interactive Sketch: Line of Best Fit

Section 3: Testing Predictions

So we have the equation for a linear model: now what? A predictive model is no good if its predictions don't actually match reality. There are a few ways of checking how well a model makes predictions, but for this activity, we'll be using something called the "sum of squared residuals (or SSR)" to be able to compare different model lines to each other.

What is a "residual"?

In regression, a residual is the difference between an observed value, like in our data set, and the value that your model predicts. It's a measure of how close or how far off your prediction is. The closer your prediction is to the observed value from your data, the smaller the residual will be. It can be calculated as:

(Residual) = (Observed value) - (Predicted value)

The sum of squared residuals approach adds up the squared residuals (i.e. residual × residual) calculated for each data point. Given two models, a lower sum of squared residuals for one model means that model's predictions are closer to the observed values and thus that model is a better model for making predictions.

Using the model above,

(ŷ: Predicted bone mineral density loss) = 2.56 × (x: Months in space) + 0

Calculate the squared residuals for the remaining data points and then calculate the sum of squared residuals.

Mission months	Loss of bone mineral density	Predicted loss of bone mineral density based on model	Residual (Observed - predicted)	Squared residual
0	0.0	0.0	0.0	0.0
1	3.2	2.6	0.6	0.4
2	6.4	5.1	1.3	1.7
3	8.4	7.7	0.7	0.5
4	9.6	10.2	-0.6	0.4
5	12.4	12.8	-0.4	0.2
6	15.4	15.4	0.0	0.0
7	18.2	17.9	0.3	0.1
8	21.5	20.5	1.0	1.0
9	23.8	23.0	0.8	0.6
10	26.3			
11	29.0			
12	30.7			

To calculate the sum of squared residuals, add up all the values in the filled-in "squared residuals" column. Unfortunately, this number is not very useful by itself. This means that the calculations in the above process need to occur over and over until the best model (i.e., the values for slope and b that result in the smallest sum of squared residuals) is found.

Interactive Sketch: Line of Best Fit

If you return to the interactive sketch, the “SSR'' readout tells you the sum of squared residuals for a given line. The yellow lines between the data points and the green fit line are the residuals for each datapoint: you’re trying to make the total of each of these (squared) the smallest number that you can.

1. Explore the sketch for a few minutes.
2. What’s the smallest SSR that you can find? What is the equation for that model?
3. Use the best model you can find to predict the amount of bone mineral density loss at 18 months and 24 months. How do these predictions compare with your earlier estimates?

Reflection & Debrief

Like decision trees, regression is another example of a potential application of machine learning. Having completed the activities above, reflect on the tasks and processes while discussing the following questions { in small groups / as a large group}:

1. How can machine learning be used with regression analysis?
    a. Prediction models based on regression analysis have parameters (such as slope and y-intercept, but also others, for non-linear models) that can be tuned to improve the accuracy of a model.
    b. By setting a goal, such as finding the smallest sum of squared residuals, machine learning can be used to find the best parameters for a given dataset and also be able to react to and recalculate if additional data is added to the dataset
2. Why would regression models be well suited to machine learning?
    a. A lot of calculations are required to find the best values for the model parameters. Computers are much faster at making those calculations than humans are, which is an important factor as the amount of data, or the number of variables, increases.

3. What affects how good our model's predictions are?
    a. Since our model's parameters were based off of the data in the dataset, we're trusting that that data is representative, i.e. that the data observed would be typical for the average astronaut. If the data isn't representative, the prediction model won't work well outside of that group of astronauts.
    b. While the data that has been collected appears to be linear for the period of 0 to 12 months, it isn't necessarily true that this will hold at the 18 or 24 month mark. The true relationship between the independent variable (months in space) and dependent variable (amount of bone mineral density loss) might take a different shape, for example, some sort of polynomial or a logistic curve.
4. How can we improve our model's predictions?
    a. More data would help to improve our model's predictions because we could have more confidence that the data that we are using to train our model parameters is representative of the actual effect of the independent variable instead of only being applicable to the astronauts in our sample.

While the data in the table above isn't real experimental data, it is based upon actual research done on astronauts. Did you know that astronauts lose an average of 1 to 2% of their bone mass for each month they spend in microgravity (e.g. space)? Visit the Canadian Space Agency's website to learn more: *https://www.asc-csa.gc.ca/eng/sciences/osm/bones.asp*

Delivery Adaptations

How might you adapt the time, space, materials, group sizes, or instructions to make this activity more approachable or more challenging? **Modifications** are ways to make the activity more accessible, **extensions** are ways to make the activity last longer or more challenging.

Modifications

- To reduce the complexity of this activity, skip over sections requiring calculations done by hand, and focus on the interactive chart models.
- For more interactive data analysis, have participants create their own charts in Google Sheets. How is data represented in different charts? Which charts might make it easier for prediction models to analyze?

Extensions

- Have participants find or create their own set of data comparing two variables, for example, leg length vs. running speed or height vs. wingspan. Have participants find or develop a dataset, and replace the data used in the interactive sketch with their own data. Use the interactive sketch to find the best model that describes their data (lowest SSR).
- Have participants develop their own questions about the activity data or their own data. For example, “What would the bone mineral density loss be at 13.5 months?” or “What would the wingspan be of someone who is 7 ft tall?”

References & Gratitude

Several activity pieces were built using p5.js, an online javascript library. It can be found at https://p5js.org/.

Canadian Space Agency (CSA, 2006). What happens to bones in space? https://www.asc-csa.gc.ca/eng/sciences/osm/bones.asp

Terms of Use

Prior to using this activity or parts thereof, you agree and understand that:

- It is your responsibility to review all aspects of this activity and ensure safety measures are in place for the protection of all involved parties.
- Any safety precautions contained in the “Safety Considerations” section of this write-up are not intended as a complete list or to replace your own safety review process.
- Actua shall not be responsible or liable for any damage that may occur due to your use of this content.
- This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. For more information, please see https://creativecommons.org/licenses/by-nc-sa/4.0/.
- You may adapt the content for your program (remix, transform, and build upon the material), providing appropriate credit to Actua and indicating if changes were made. No sharing of content with third parties without written permission from Actua.

About Actua

Actua is Canada’s leading science, technology, engineering and mathematics (STEM) youth outreach network, representing a growing network of over 40 universities and colleges across the country. Each year 350,000 young Canadians in over 500 communities nationwide are inspired through hands-on educational workshops, camps and community outreach initiatives. Actua focuses on the engagement of underrepresented youth through specialized programs for Indigenous youth, girls and young women, at-risk youth and youth living in Northern and remote communities. For more information, please visit us online at www.actua.ca and on social media: Twitter, Facebook, Instagram and YouTube!

Appendices

Appendix A: Career & Mentor Connections

- Machine learning researcher
- Programmer
- Software Engineer
- Researcher