Leveraging the National Research Platform to build a Scalable Research and Education Environment

Tyler Pearson – Director, Informatics
Mark Laufersweiler, Ph.D. – Research Data Specialist
Recognizing the Giants

• National (Pacific) Research Platform and their partners
• Great Plains Network and former Executive Director James Deaton
About Nautilus

- NSF FUNDED 6 GRANTS TOTALING OVER $27,000,000
- 50+ PARTNERING INSTITUTIONS
- RUNS ON KUBERNETES AND PARTNER PROVIDED HARDWARE
- COMMUNITY SHARED RESOURCES, GITLAB, ONLINE CHAT, WORKSHOPS
• An open-source system for automating deployment, scaling, and management of containerized applications
• A multi-user version of the notebook designed for companies, classrooms and research labs
OUr Issues

- Installation of Software on Local Machines
- Research Reproducibility
- Lack of a Common Coding Environment
- Variety of User Hardware
Software Installation Issues

- Varying User Environments
- Policy Restrictions
- Inconsistent Versioning
- Conflicting Software
- Reproducibility Challenges
Hardware Issues

- Chromebooks
- Legacy Laptops
- Apple's M1 and M2 Architecture
- Research
  - Nvidia GPU Model Availability
OUr Deployment

- Created initial Nautilus namespace for testing
- Onboarded first instructor alpha tester from meteorology
- Determined metrics and how to collect
- Speed bumps and how to resolve live issues
Phase 1 Pilot Timeline

Dec. 2021
Deployed our first JupyterHub in Nautilus

Jan. 2022
Testing and learning capabilities
Documentation and infrastructure scripting

Feb. 2022
Presented at ATE which led to our first custom container for METR 3334

Mar. 2022
Hosted first Software Carpentry Workshop

Apr. 2022
Shared our deployment configuration for a KU Library workshop

Jun. 2022
Provided to Meteorology REU summer research project with an NLP environment

Aug. 2022
Implemented metrics gathering

Sep. 2022
Hosted first class with custom software and libraries – METR 1313

Oct. 2022
Expanded pilot to all METR 1313 sections

Nov. 2022

Dec. 2022

Jan. 2023

Our Custom Environments in JupyterHub
R and RStudio Stack
RStudio Running
OpenRefine Running
Meteorology
Custom
Environment
A step-by-step guide to linear regression in R

Published on February 05, 2020 by Rebecca Beaven. Revised on December 14, 2020.

Linear regression is a regression model that uses a straight line to describe the relationship between variables. It finds the line of best fit through your data by searching for the value of the regression coefficient(s) that minimizes the total error of the model.

There are two main types of linear regression:
- Simple linear regression uses only one independent variable
- Multiple linear regression uses two or more independent variables

In this step-by-step guide, we will walk you through linear regression in R using two sample datasets.

**Simple linear regression**

The first dataset contains observations about income (in a range of $10k to $75k) and happiness (rated on a scale of 1 to 10) in an imaginary sample of 500 people. The income values are divided by 10,000 to make the income data match the scale of the happiness scores (so a value of 83 represents $83,000, 8 is $80,000, etc.)

**Multiple linear regression**

The second dataset contains observations on the percentage of people biking to work each day, the percentage of people smoking, and the percentage of people with heart disease in an imaginary sample of 500 towns.

Download the sample datasets to try it yourself:

- Simple Regression Dataset
- Multiple Regression Dataset

**Table of contents**

- Getting started in R
  - Load the data into R
  - Make sure your data meet the assumptions
  - Perform the linear regression analysis
  - Check for homoscedasticity
  - Visualize the results with a graph
  - Report your results

**Getting started in R**
Metrics November 2022 to Present

• 2,592 Unique Sessions by 335 users
• Median Session Length 1.91 hours

<table>
<thead>
<tr>
<th>Top Environments (Pods)</th>
<th>Session Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meteorology</td>
<td>2102</td>
</tr>
<tr>
<td>Scipy</td>
<td>188</td>
</tr>
<tr>
<td>RStudio</td>
<td>63</td>
</tr>
<tr>
<td>Tensorflow</td>
<td>53</td>
</tr>
<tr>
<td>R</td>
<td>25</td>
</tr>
<tr>
<td>OpenRefine</td>
<td>16</td>
</tr>
</tbody>
</table>
Typical Class Session at Course Time
<table>
<thead>
<tr>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>library.ucla.edu</td>
</tr>
<tr>
<td>ncsu.edu</td>
</tr>
<tr>
<td>ou.edu</td>
</tr>
<tr>
<td>ouhsc.edu</td>
</tr>
<tr>
<td>ku.edu</td>
</tr>
<tr>
<td>harvard.edu</td>
</tr>
<tr>
<td>greatplains.net</td>
</tr>
<tr>
<td>kanren.net</td>
</tr>
<tr>
<td>students.niu.edu</td>
</tr>
<tr>
<td>usd.edu</td>
</tr>
<tr>
<td>vt.edu</td>
</tr>
<tr>
<td>okstate.edu</td>
</tr>
<tr>
<td>calu.edu</td>
</tr>
<tr>
<td>lsu.edu</td>
</tr>
<tr>
<td>unm.edu</td>
</tr>
<tr>
<td>miamioh.edu</td>
</tr>
<tr>
<td>udel.edu</td>
</tr>
<tr>
<td>valpo.edu</td>
</tr>
<tr>
<td>csus.edu</td>
</tr>
<tr>
<td>wichita.edu</td>
</tr>
</tbody>
</table>
What We Are Learning

Speed Bumps
We do not own the hardware and infrastructure
- Slowdowns can occur when supporting nodes/computes crash
We cannot control the network
Need of tools when diagnosing issues of learners

How we help
Realtime support
Environment Design and Build
Collect Feedback from learners and instructors
Providing feedback to Nautilus
What’s Next?

Documentation
Internal
Learners/Instructors

Scoping for what
the Library can
support

Add other
domain courses
to pilot
Digital Humanities
Biology

Nautilus
Frameworks for
Researchers

Local Partnering
Thank You and Questions?

Tyler Pearson
Director of Informatics
Univ of Oklahoma Libraries
tdpearson@ou.edu

Mark Laufersweiler, Ph.D.
Research Data Specialist
Univ of Oklahoma Libraries
laufers@ou.edu