INSTITUTE FOR DEFENSE ANALYSES

DATAWorks 2023: Development of a Wald-Type Statistical Test to Compare Live Test Data and M&S Predictions

Elliot AJ Bartis, Project Leader

Carrington A. Metts
Curtis G. Miller

April 2023

IDA Document NS D-33406

Log: H 2023-000065

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task BD-09-2299(32), "USW", for the Office of the Director, Operational Test and Evaluation.. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Dr. Kyle E. Remley and Dr. Olivia Gozdz from the Operational Evaluation Division

For more information:
Dr. Elliot AJ Bartis, Project Leader
ebartis@ida.org • (703) 845-6853

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-33406

DATAWorks 2023: Development of a Wald-Type Statistical Test to Compare Live Test Data and M&S Predictions

Elliot AJ Bartis, Project Leader

Carrington A. Metts
Curtis G. Miller

Executive Summary

This work describes the development of a statistical test created in support of ongoing verification, validation, and accreditation (VV&A) efforts for modeling and simulation (M&S) environments. The test decides between a null hypothesis of agreement between the simulation and reality, and an alternative hypothesis stating the simulation and reality do not agree. To do so, it generates a Wald-type statistic that compares the coefficients of two generalized linear models that are estimated on live test data and analogous simulated data, then determines whether any of the coefficient pairs are statistically different.

The test was applied to two logistic regression models that were estimated from live torpedo test data and simulated data from the Naval Undersea Warfare Center's (NUWC) Environment Centric Weapons Analysis Facility (ECWAF). The test did not show any significant differences between the live and simulated tests for the scenarios modeled by the ECWAF. More work is needed to fully validate the ECWAF's performance, but this finding suggests that the facility is adequately modeling the various target characteristics and environmental factors that affect in-water torpedo performance.

The primary advantage of this test is that it can handle cases where one or more variables are estimable in one model but missing or inestimable from the other. It is possible to simply create the linear models on the common set of variables, but this results in the omission of potentially useful test data.

Instead, this approach identifies the mismatched coefficients and combines them with the model's intercept term, thus allowing the user to consider models that are created on the entire set of available data. Furthermore, the test was developed in a generalized manner without any references to a specific dataset or system. Therefore, other researchers who are conducting VV&A processes on other operational systems may benefit from using this test for their own purposes.

DATAWorks 2023: "Development of a Wald-Type Statistical Test to Compare Live Test Data and M&S Predictions"

Carrington Metts

Curtis Miller

Project Leader: Elliot Bartis

2/24/2023

Institute for Defense Analyses

730 East Glebe Road ● Alexandria, Virginia 22305

What methods can researchers use to ensure M&S predictions agree with live test data? How can this methodology be standardized?

Environment Centric Weapons Analysis Facility (ECWAF)

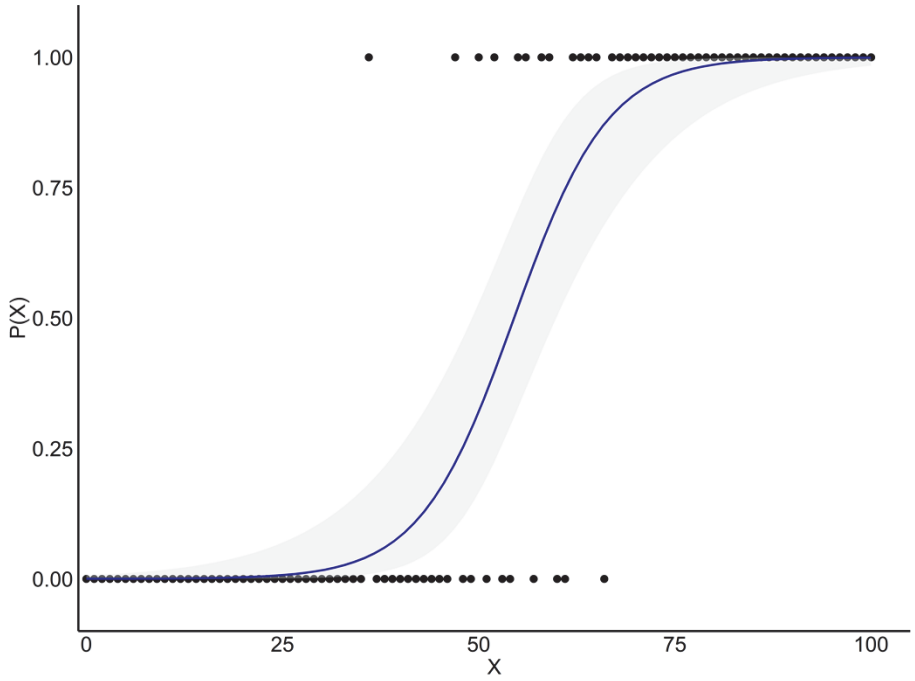
- NUWC-operated MK 48 simulation facility
- Real-time, hardware-in-the-loop
- Advantages
  - More cost-effective
  - Test novel environmental and adversarial conditions
- Risks
  - ECWAF predictions may not be correct
- VV&A is ongoing

NUWC – Naval Undersea Warfare Center; VV&A – Verification, Validation, & Accreditation

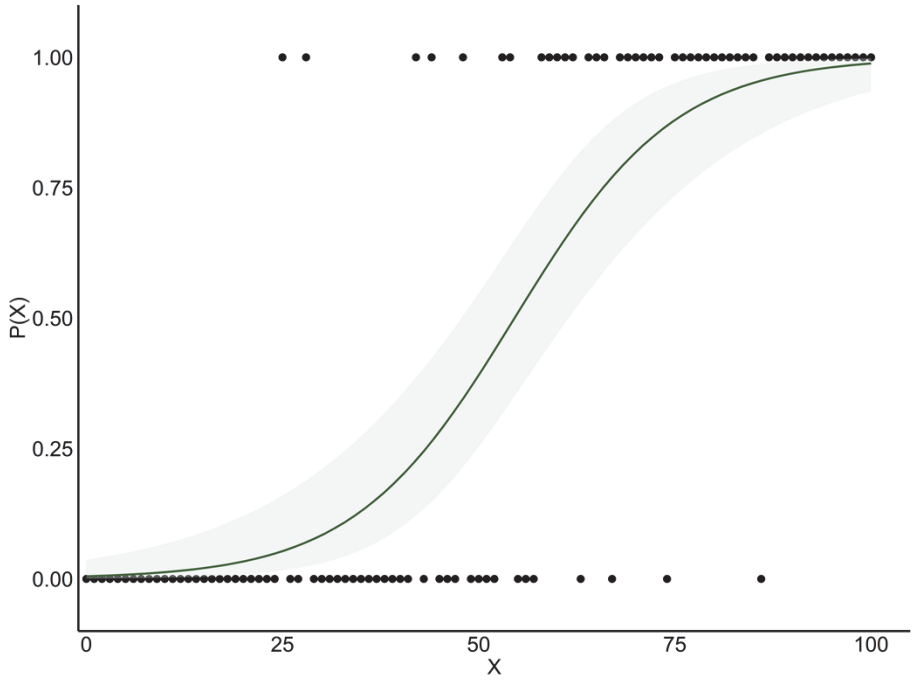
Image source: https://www.lockheedmartin.com/en-us/products/mk-48-mod-7-common-broadband-advanced-sonar-system-cbass-heavyweight-torpedo.html

Generalized linear models (GLMs) are used to evaluate the ECWAF's performance

Logistic Regression for Live Test Data*



Logistic Regression for ECWAF Data*



**Data are fully simulated and not intended to reflect any results from the ECWAF or live tests. Plots are for illustrative purposes only.*

A Wald-type hypothesis test may be used to compare in-water and live test data

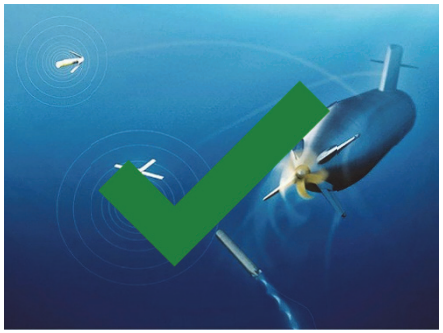
$$W = \left(\hat{\beta}_{sim} - \hat{\beta}_{live}\right)' V^{-1} \left(\hat{\beta}_{sim} - \hat{\beta}_{live}\right)$$

H_0 (null): Simulations and live data agree

H_A (alternative): Simulations and live data do not agree

- V is the covariance matrix
- W approximately follows a χ^2 distribution

Problems arise if variables are estimable in one model and not the other



$$g\left(E\left[Y_j^{live}\right]\right) = \beta_0 + \beta^{depth} \mathrm{X}_j^{depth}$$

$$g\left(E\left[Y_j^{sim}\right]\right) = \beta_0 + \boxed{\beta^{CM} \mathrm{X}_j^{\mathrm{CM}}} + \beta^{depth} \mathrm{X}_j^{depth}$$

The Wald test requires all coefficients to be matched. How can we handle mismatched coefficients?

Image source: https://www.thedrive.com/the-war-zone/33467/the-shadowy-world-of-submarine-and-ship-launched-torpedo-countermeasures-an-explainer

Generalized linear models provide a framework for addressing mismatched coefficients

$$g\left(E\left[Y_j^{live}\right]\right) = \beta_0 + \beta^{depth} \mathrm{X}_j^{depth}$$

With countermeasures
Depth of 0

+

Depth adjustment

Generalized linear models provide a framework for addressing mismatched coefficients

$$g\left(E\left[Y_j^{live}\right]\right) = \beta_0 + \beta^{depth} \mathrm{X}_j^{depth}$$

With countermeasures Depth of 0 + Depth adjustment

$$g\left(E\left[Y_j^{sim}\right]\right) = \beta_0 + \beta^{CM} \mathrm{X}_j^{\mathrm{CM}} + \beta^{depth} \mathrm{X}_j^{depth}$$

No countermeasures Depth of 0 + Countermeasure adjustment + Depth adjustment

Generalized linear models provide a framework for addressing mismatched coefficients

$$g\left(E\left[Y_j^{live}\right]\right) = \beta_0 + \beta^{depth} \mathrm{X}_j^{depth}$$

With countermeasures Depth of 0 + Depth adjustment

$$g\left(E\left[Y_j^{sim}\right]\right) = \beta_0 + \beta^{CM} \mathrm{X}_j^{\mathrm{CM}} + \beta^{depth} \mathrm{X}_j^{depth}$$

No countermeasures Depth of 0 + Countermeasure adjustment + Depth adjustment

With countermeasures; Depth of 0

Generalized linear models provide a framework for addressing mismatched coefficients and performing a Wald test.

1. Identify all mismatched coefficients
2. Find fixed value for corresponding variable
3. Multiply mismatched coefficient by fixed value
4. Add result to intercept term
5. Adjust covariance matrix

Result is an R function that automatically detects and addresses mismatched coefficients

```
> result <- wald_coef_compare(model1, model2)
> result

        Wald Type Test for Differences in Coefficients

data:  model1 and model2
statistic = 12.64, df = 7, p-value = 0.08139
alternative hypothesis: The two models are not identical
```

Results are generated from unrelated, publicly available data and do not reflect the ECWAF's performance.

When applied to ECWAF data, the test does not find meaningful differences between M&S and live test data.

Development of a Wald-Type Statistical Test to Compare Live Test Data and M&S Predictions

Carrington Metts

with Curtis Miller and Elliot Bartis

Institute for Defense Analyses

Problem Statement

As part of the verification, validation, and accreditation (VV&A) process for the Environment Centric Weapons Analysis Facility (ECWAF), researchers must determine whether the facility's predictions agree with in-water torpedo test data. Even if the final predictions are statistically similar, do the simulations correctly model the effects of environmental conditions?

Project Goal

To create a robust statistical test in R that can be used to determine if the output of M&S agrees with live test data.

A satisfactory statistical test should be:

1. *Generalizable*: The test should be able to be applied to any combination of live test data and M&S predictions.
2. *Interpretable*: The test's output should align with existing R functions, such as *t.test*.
3. *Robust*: The test should handle instances where environmental variables are present in one case (live test or M&S), but absent in the other.

About the ECWAF

The ECWAF is a real-time, hardware-in-the-loop simulation facility operated by the Naval Undersea Warfare Center (NUWC) to assist in MK 48 torpedo operational testing. Within the ECWAF, a torpedo processor is connected to inputs that mimic live test conditions. Data from the processors indicates whether the torpedo successfully struck its target.

The ECWAF allows for more cost-effective testing under a wider range of environmental and adversarial conditions, compared to traditional in-water test plans. However, the VV&A process is ongoing.

Evaluating ECWAF Results

As part of the VV&A process for the ECWAF, IDA researchers fit generalized linear models (GLMs) to the available set of live test data and an analogous set of ECWAF simulations. Both models describe the torpedo's effectiveness as a function of the target's and scenario's characteristics.

Generalized linear models take the following form:

$$g(E[Y_j]) = \beta_0 + \sum_{i=1}^{k} \beta_i X_{ij}$$

Here, X_{ij} represents a predictor for a given observation j. The intercept and model coefficients are represented by β_0 and $\beta_1,\ldots,\beta_k$, respectively. The quantity $g(E[Y_j])$ represents the expected value of the response variable, where g is a function that maps the allowed values of the response variable to the real number line.

After creating the two models, we created a function to compare each pair of corresponding coefficients. If none of the pairs of coefficients are statistically different, that suggests that the simulations are accurately representing the real-world environmental conditions.

$$g(E[Y_j^{live}]) = \beta_0 + \beta^{depth} X_j^{depth}$$

With countermeasures Depth of 0 + Depth adjustment

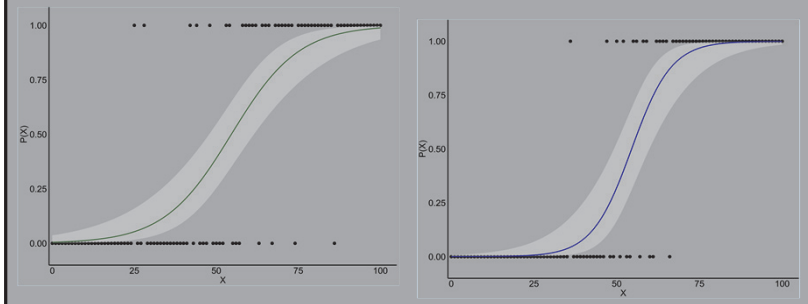
Mismatched coefficients can be combined with the model's intercept to enable a pairwise comparison.

$$g(E[Y_j^{sim}]) = \beta_0 + \beta^{CM} X_j^{CM} + \beta^{depth} X_j^{depth}$$

No countermeasures Depth of 0 + Countermeasure adjustment + Depth adjustment

With countermeasures; Depth of 0

Are these generalized linear models describing the same underlying data? What statistical techniques can we use to determine the answer?



Data are fully simulated and not intended to reflect any results from the ECWAF or live tests. Plots are for illustrative purposes only.

The result is a flexible, easy-to-use statistical test that computes the value of the Wald statistic and the p-value for any two generalized linear models.

```
> result <- wald_coef_compare(model1, model2)
> result

        Wald Type Test for Differences in Coefficients

data:  model1 and model2
statistic = 12.64, df = 7, p-value = 0.08139
alternative hypothesis: The two models are not identical
```

Results are displayed for illustrative purposes only. Data and results do not reflect any outcome from the ECWAF or live tests.

Wald Type Tests

Wald tests are a type of hypothesis test designed to simultaneously test multiple hypotheses. They determine whether a set of observations agrees with what is expected by calculating the distance between each observation and its expected value, normalized by the variance.

For this work, a Wald test was used to compare each coefficient from the linear model with its in-water analogue:

$$W = (\hat{\beta}_{sim} - \hat{\beta}_{live})' V^{-1} (\hat{\beta}_{sim} - \hat{\beta}_{live})$$

The Wald statistic assumes a null and alternative hypothesis:

H_0 (null): M&S outputs agree with live test observations

H_A (alternative): M&S outputs do not agree with live test observations

The test statistic approximately follows a χ^2 distribution, which allows for p values to be determined.

Handling Missing Coefficients

The Wald test can only be applied if every coefficient in the simulation's GLM has an analogue in the live test data. In many operational scenarios, this will not be the case.

For example, consider a scenario where live tests are conducted on a target that employs countermeasures. Corresponding ECWAF simulations are conducted both with and without countermeasures. The linear model for the live tests will not have a coefficient for countermeasures, but the ECWAF model will.

Generalized linear models provide a framework for addressing mis-matched coefficients. A linear model's intercept term (β_0) represents the model's predictions when all variables are equal to 0. The other terms (denoted as $\beta_i X_{ij}$, where i indicates the variable and j represents an observation) represent the change in this baseline prediction.

We included functionality in our code to address mismatched coefficients:

1. Identify coefficients that are present in one GLM but not the other.
2. For each mismatched coefficient, identify the corresponding variable's fixed value in the other model (and default to 0, if it is not present).
3. Multiply the fixed value by the mismatched coefficient and add the result to the intercept term.
4. For the model's covariance matrix, multiply the corresponding row by the fixed value and add the row to the matrix's intercept row.

Results

The result is an informal R package that can be applied to any two GLMs with a single line of code. The function's output mirrors the format of other popular statistical functions, such as *t.test*.

When applied to ECWAF and corresponding live test data, the function did not detect any significant differences between the two linear models. This finding is evidence that the ECWAF is performing well for the range of environmental conditions that were tested. However, more VV&A work is needed before its predictions can be fully trusted.

Conclusions

This work describes the development of a robust statistical test that compares the coefficients of two linear models. Other researchers who are conducting VV&A analyses on M&S outputs may benefit from using the test to ensure their simulations are accurately modeling a range of real-world environmental conditions.

Acknowledgments

Thank you to Curtis Miller for his contributions through the entire test development process.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-04-2023	Final	Apr 2023

4. TITLE AND SUBTITLE

DATAWorks 2023 Development of a Wald-Type Statistical Test to Compare Live Test Data and M&S Predictions

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Ms. Carrington A. Metts
Dr. Curtis G. Miller

5d. PROJECT NUMBER

BD-09-2299(32)

5e. TASK NUMBER

229932

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER

NS D-33406

H 2023-000065

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington D.C. 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

Project leader: Dr. Elliot AJ Bartis

14. ABSTRACT

This work describes the development of a statistical test created in support of ongoing verification, validation, and accreditation (VV&A) efforts for modeling and simulation (M&S) environments. The test decides between a null hypothesis of agreement between the simulation and reality, and an alternative hypothesis stating the simulation and reality do not agree. To do so, it generates a Wald-type statistic that compares the coefficients of two generalized linear models that are estimated on live test data and analogous simulated data, then determines whether any of the coefficient pairs are statistically different.

The test was applied to two logistic regression models that were estimated from live torpedo test data and simulated data from the Naval Undersea Warfare Center's (NUWC) Environment Centric Weapons Analysis Facility (ECWAF). The test did not show any significant differences between the live and simulated tests for the scenarios modeled by the ECWAF. While more work is needed to fully validate the ECWAF's performance, this finding suggests that the facility is adequately modeling the various target characteristics and environmental factors that affect in-water torpedo performance.

The primary advantage of this test is that it is capable of handling cases where one or more variables are estimable in one model but missing or inestimable from the other. While it is possible to simply create the linear models on the common set of variables, this results in the omission of potentially useful test data. Instead, this approach identifies the mismatched coefficients and combines them with the model's intercept term, thus allowing the user to consider models that are created on the entire set of available data. Furthermore, the test was developed in a generalized manner without any references to a specific dataset or system. Therefore, other researchers who are conducting VV&A processes on other operational systems may benefit from using this test for their own purposes.

15. SUBJECT TERMS

USW; VV&A; M&S; Undersea warfare; Torpedoes; ECWAF; Wald test

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Elliot AJ Bartis
Unclassified	Unclassified	Unclassified	SAR	16	19b. TELEPHONE NUMBER *(Include area code)* 703-845-6853