INSTITUTE FOR DEFENSE ANALYSES

# Introduction to ciTools

Laura Freeman, *Project Leader*
John Haman
Matthew Avery

August 2017



IDA Document NS D-8702

Log: 2017-000501

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

*The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.*

About This Publication

ciTools is an R package for working with model uncertainty. It gives users access to confidence and prediction intervals for the fitted values of (log-) linear models, generalized linear models, and (log-) linear mixed models. Additionally, ciTools provides functions to determine probabilities and quantiles of the conditional response distribution given each of these models. This briefing introduces the package and provides simple illustrations for using ciTools to perform inference and plot results.

For more information:
Laura J. Freeman, Project Leader
lfreeman@ida.org • (703) 845-2084

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-8702

# Introduction to ciTools

Laura Freeman, *Project Leader*
John Haman
Matthew Avery

# Executive Summary

## A. Quantifying Uncertainty

Analysts at IDA developed a new R package, ciTools, to quantify uncertainty easily and quickly, addressing a gap in the software.

Quantifying uncertainty is a critical part of good quantitative data analysis. IDA analysts frequently use statistical models to estimate system performance and use these estimates in reports. Estimates without uncertainty boundaries can be misleading, yet many software packages do not provide straightforward ways to compute intervals around estimated values. Confidence intervals quantify the uncertainty of an estimate of an average, such as the average target location error (TLE) for a sensor. Prediction intervals provide a range within which we would expect a new individual observation to fall, using estimated population variance. Rather than bound the average TLE for a sensor, a prediction interval bounds the TLE of an individual target location provided by the sensor. Prediction intervals are often more intuitive and informative to operators who want to know what to expect when they use a system in the field. Similarly, probability estimates answer the question, "What is the probability that the sensor will provide me with a target location that is at least $Q$ meters away from the target?", while quantile estimates provide a bound for a given percentage of all individual observations. For example a Circular Error Probable of 90 is equivalent to the 0.9 quantile of a system's TLE. Confidence intervals, prediction intervals, probabilities, and quantiles all answer important questions that come up frequently in data analysis done at IDA.

R is a popular, free, open-source statistical software program used by many IDA analysts; unfortunately, most statistical modeling packages in R do not make it convenient to get uncertainty estimates. Some modeling packages in R make confidence intervals available, but focus is typically on model parameters rather than estimates of the response variable such as a measure of system performance. Prediction intervals, quantiles, and probability estimates are rarely available, if at all, and require analysts to write additional code. The syntax for generating these estimates can vary substantially depending on the type of statistical model used. The package ciTools addresses

these problems by presenting a convenient, consistent set of functions for estimating response variable uncertainty across a wide range of common statistical models used by IDA analysts.

## B. Attributes of ciTools

The R package ciTools provides a uniform, model-invariant, data-first approach for working with model uncertainty in R. Table 1 shows the different types of data for which ciTools provides functionality. For each kind of data, a different family of statistical model is used, and an associated different process is required to accurately estimate uncertainty. In many cases, these processes are difficult and require careful thought to do correctly. By automating the process and making syntax consistent for different types of data, ciTools makes it easier for data analysts to quickly provide accurate results.

The four main functions in ciTools have uniform syntax, making it easy for analysts to easily switch from one type of uncertainty estimate to another. The functions look the same, users provide inputs in the same order for each function, and defaults are consistent across each function. Figure 1 shows the four main functions and their uniform syntax.

Functions in ciTools are also model-invariant, meaning that analysts only have to learn how to do things once. Regardless of the type of data and statistical model the analyst is working with, the same code is used to generate a confidence interval or quantile estimate. This makes it easy for analysts who have used ciTools for one type of data analysis to use it again for another type of analysis.

**Table 1. Scope of ciTools**

| | Confidence Intervals | Prediction Intervals | Probabilities | Quantiles |
|---|---|---|---|---|
| **Linear Data** | ✓ | ✓ | ✓ | ✓ |
| **Count/ Binary Data** | ✓ | ✓ | ✓ | ✓ |
| **Skewed Data** | ✓ | ✓ | ✓ | ✓ |
| **Random Group Data** | ✓ | ✓ | ✓ | ✓ |
| **Reliability Data** | Future Work | Future Work | Future Work | Future Work |
| **Skewed, Random Group Data** | In Progress | ✓ | ✓ | ✓ |

| Confidence Intervals | Prediction Intervals |
|---|---|
| `add_ci(data, model, …)` | `add_pi(data, model, …)` |
| **Probabilities** | **Quantiles** |
| `add_probs(data, model, quantile, …)` | `add_quantile(data, model, probability, …)` |

**Figure 1. The four main functions in ciTools have consistent look and syntax.**

Finally, ciTools uses a data-first approach, which makes it fit seamlessly into the "tidyverse", a popular set of R packages used for cleaning, processing, and plotting data. A data table is the first argument of each function, and each function returns that same data table augmented with the requested interval, probability, or quantile estimate. This approach makes it easy for analysts to write legible code and is consistent with reproducible research best practices. It also makes plotting these uncertainty estimates straightforward because the factors used to generate the model predictions and uncertainty estimates remain in the data table.

## C. Summary

At present, ciTools is available to IDA analysts and the public a large on both the Comprehensive R Archive Network and GitHub. A full set of help files and multiple tutorials are accessible through R.


# `ciTools`

A New **R** Package to Ease Model Inference

John Haman

Matthew Avery

Interval estimates are difficult but vital in OED research.

`ciTools` makes working with them **easy.**

# Outline

1. Motivating Examples

2. Design of `ciTools`
    1. Uniformity
    2. Model Invariance
    3. “Data First” Approach

3. Example: `ciTools` and Graphics

4. Scope of `ciTools` and Learning More

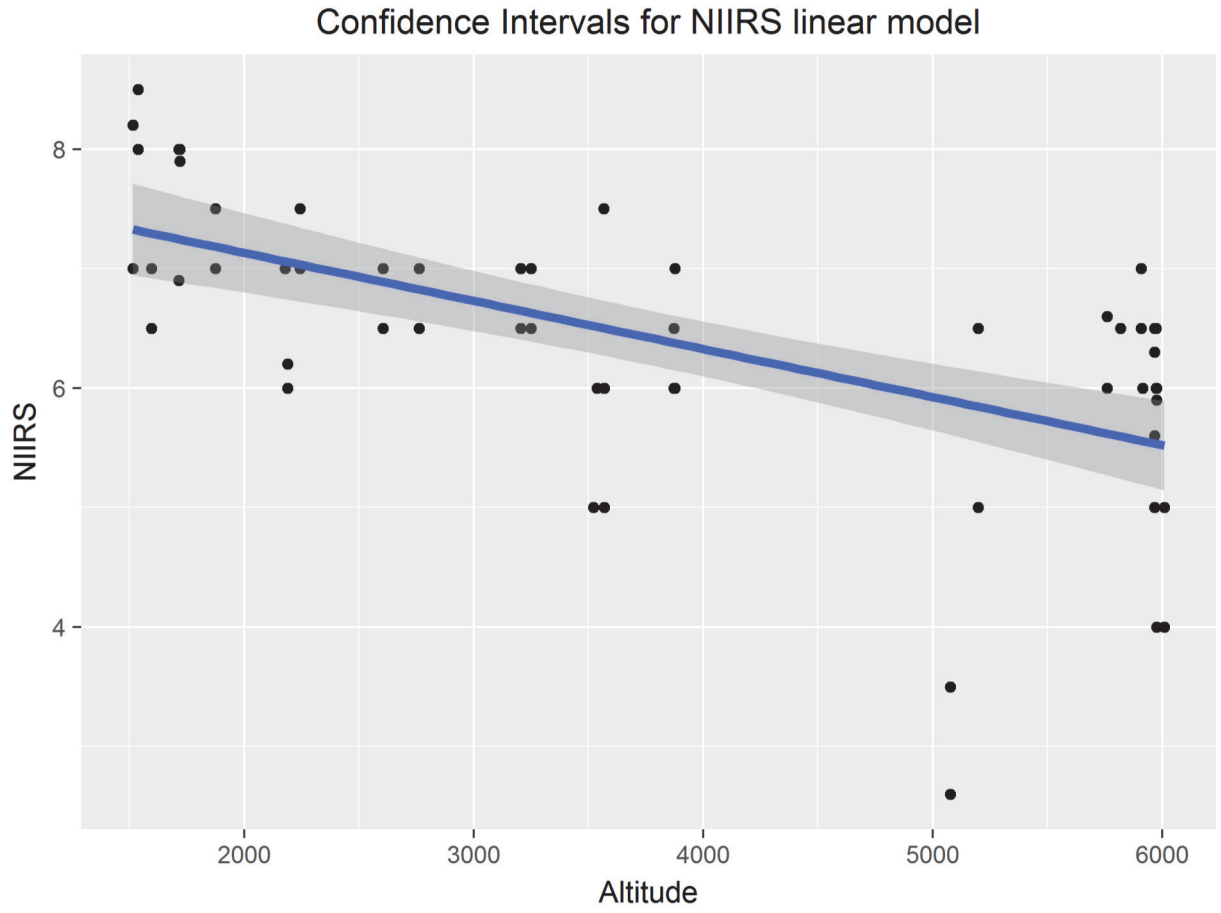
# Unmanned Reconnaissance Aircraft with Imagery Payload

NIIRS = 6


NIIRS = 8

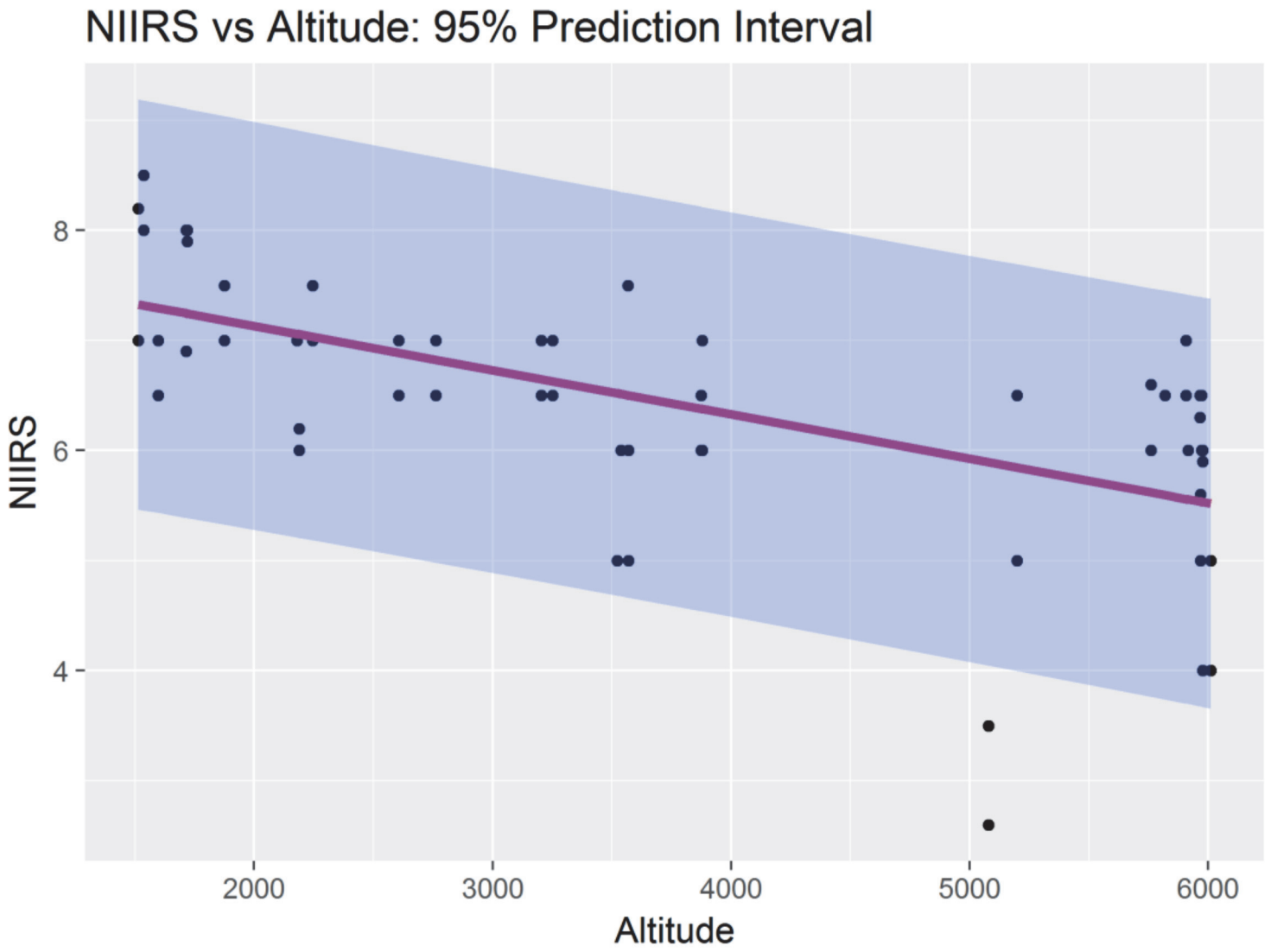
Image Source: fas.org

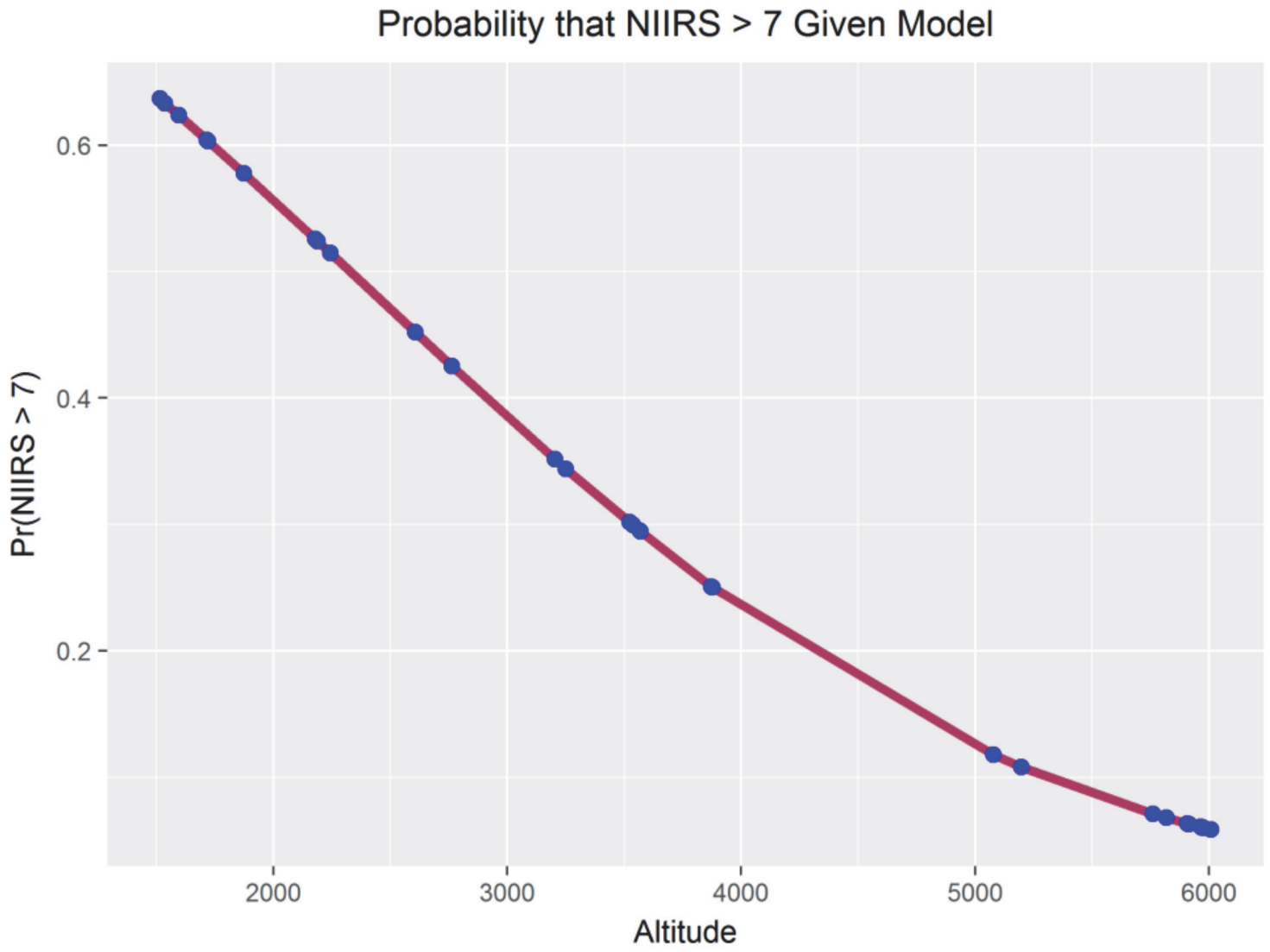
# Confidence intervals for simple linear models are easy and convenient.




Possible to get confidence intervals *while* plotting in **R**

# Doing other things can be more difficult …




Probability that NIIRS > 7 Given Model
Pr(NIIRS > 7)
0.6
0.4
0.2
2000
3000
4000
5000
6000
Altitude

`ciTools` presents a uniform, model-invariant, "data first" approach to working with model uncertainty in R.

# Scope

Functionality

Data types

| | Confidence Intervals | Prediction Intervals | Probabilities | Quantiles |
|---|---|---|---|---|
| Linear Data | | | | |
| Count/Binary Data | | | | |
| Skewed Data | | | | |
| Random Group Data | | | | |
| Reliability Data | | | | |
| Skewed, Random Group Data | | | | |

# Minimal, Uniform Design

# Uniformity in `ciTools`

`ciTools` has many capabilities,
and they all use the same syntax:

| **Confidence Intervals** | **Prediction Intervals** |
|---|---|
| `add_ci(data, model, …)` | `add_pi(data, model, …)` |
| **Probabilities** | **Quantiles** |
| `add_probs(data, model, quantile, …)` | `add_quantile(data, model, probability, …)` |

# NIIRS Data: Generating confidence intervals

```
> add_ci(dat, fit)
# A tibble: 65 x 5
   NIIRS Altitude     pred LCB0.025 UCB0.975
   <dbl>    <int>    <dbl>    <dbl>    <dbl>
 1   8.0     1537 7.318570 6.938266 7.698875
 2   7.0     1874 7.182952 6.837865 7.528039
 3   7.0     3251 6.628808 6.389208 6.868409
 4   6.5     2606 6.888375 6.609529 7.167220
 5   7.5     3568 6.501239 6.271181 6.731296
 6   6.5     5818 5.595775 5.241334 5.950216
 7   7.0     5908 5.559557 5.195761 5.923352
 8   6.3     5966 5.536216 5.166307 5.906125
 9   6.5     1597 7.294425 6.920550 7.668299
10   8.0     1716 7.246536 6.885218 7.607853
# ... with 55 more rows
```

# NIIRS Data: Generating probability estimates

```
> add_probs(dat, fit, q = 7)
# A tibble: 65 x 4
   NIIRS Altitude     pred prob_less_than7
   <dbl>    <int>    <dbl>           <dbl>
 1   8.0     1537 7.318570       0.3667974
 2   7.0     1874 7.182952       0.4222097
 3   7.0     3251 6.628808       0.6560007
 4   6.5     2606 6.888375       0.5479452
 5   7.5     3568 6.501239       0.7052830
 6   6.5     5818 5.595775       0.9320821
 7   7.0     5908 5.559557       0.9367160
 8   6.3     5966 5.536216       0.9395578
 9   6.5     1597 7.294425       0.3764595
10   8.0     1716 7.246536       0.3958999
# ... with 55 more rows
```

# Model Invariance

# Model-Invariant Design

The same functions with the same syntax can be applied to each type of model supported.

# Generating a prediction interval for a simple linear model without `ciTools`

```
## Linear Model Prediction Interval
predict(model, SE = TRUE, interval = "prediction")
```

# Generating a prediction interval for a poisson GLM without `ciTools`

```r
## Poisson Model Prediction Interval
## Design your own simulation :(
sims <- 1000
obs <- NROW(tb)
modmat <- model.matrix(fit)
response_distr <- fit$family$family
inverselink <- fit$family$linkinv
out <- inverselink(predict(fit, tb))
sims <- arm::sim(fit, n.sims = nSims)
sim_response <- matrix(0, ncol = nSims, nrow = nPreds)

for (i in 1:nPreds){
  if(response_distr == "poisson"){
    sim_response[i,] <- rpois(n = nSims,
    lambda = inverselink(rnorm(nPreds,sims@coef[i,]
    %*% modmat[i,], sd = sims@sigma[i])))
  }
}

lwr <- apply(sim_response, 1,
             FUN = quantile, probs = alpha / 2, type = 1)
upr <- apply(sim_response, 1,
             FUN = quantile, probs = 1 - alpha / 2, type = 1)

data[["pred"]] <- out
data[[names[1]]] <- lwr
data[[names[2]]] <- upr
```

## `ciTools` allows you to use the same code for the inference you want even when you fit models of different types.

```
## Same one-line command
add_pi(dat, my_model)
add_pi(dat, my_other_model)
add_pi(dat, yet_another_model)
```

# “Data First” Syntax

The first input to all functions is a data table, and the only output is the same data table augmented by what you asked for.

# Why “Data First”?

1. Facilitates function “chains” and plotting through use of the “pipe” operator.

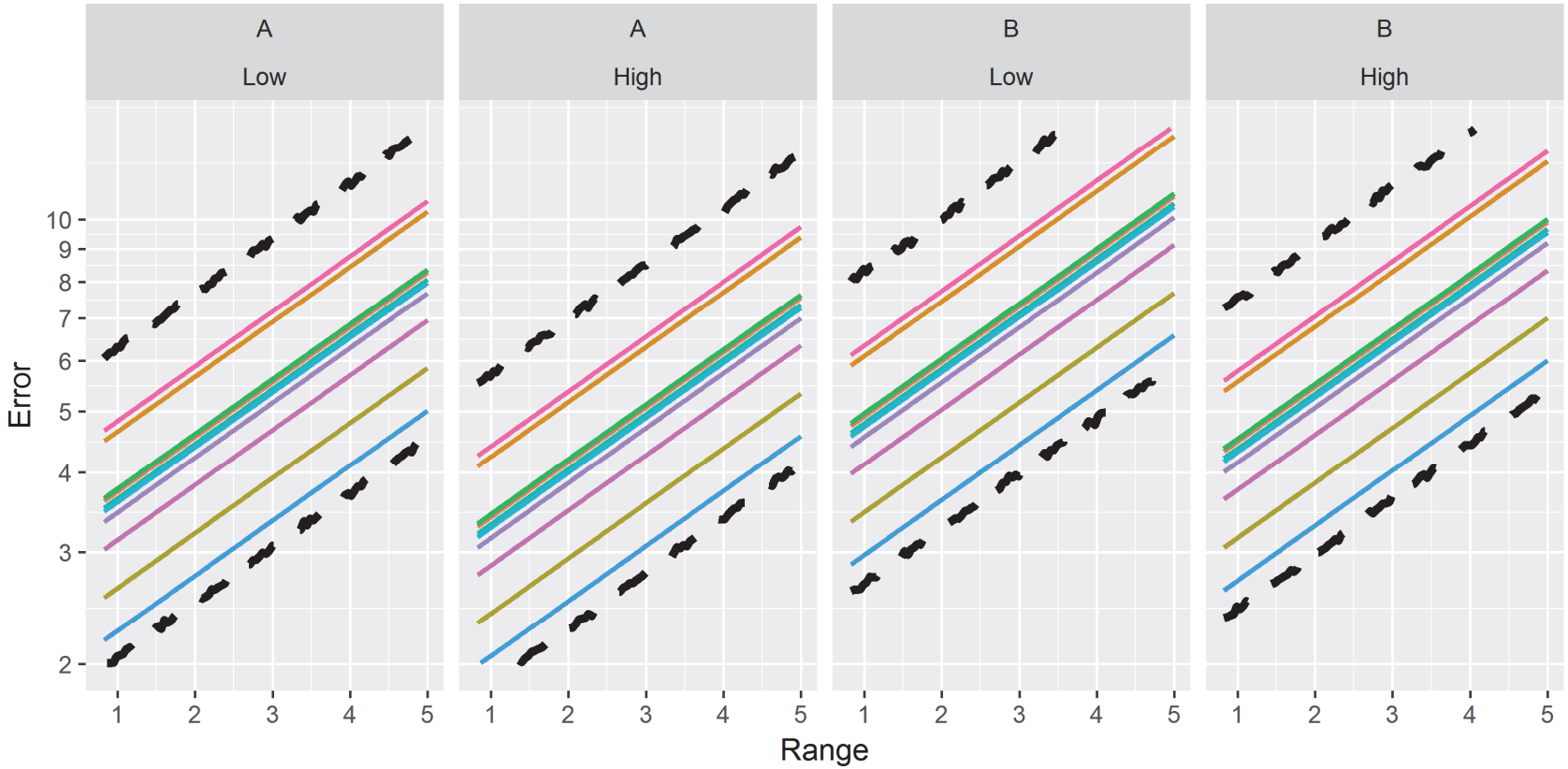
2. Improves code legibility.

3. Consistency with the `tidyverse`.

# Plotting with `ciTools`

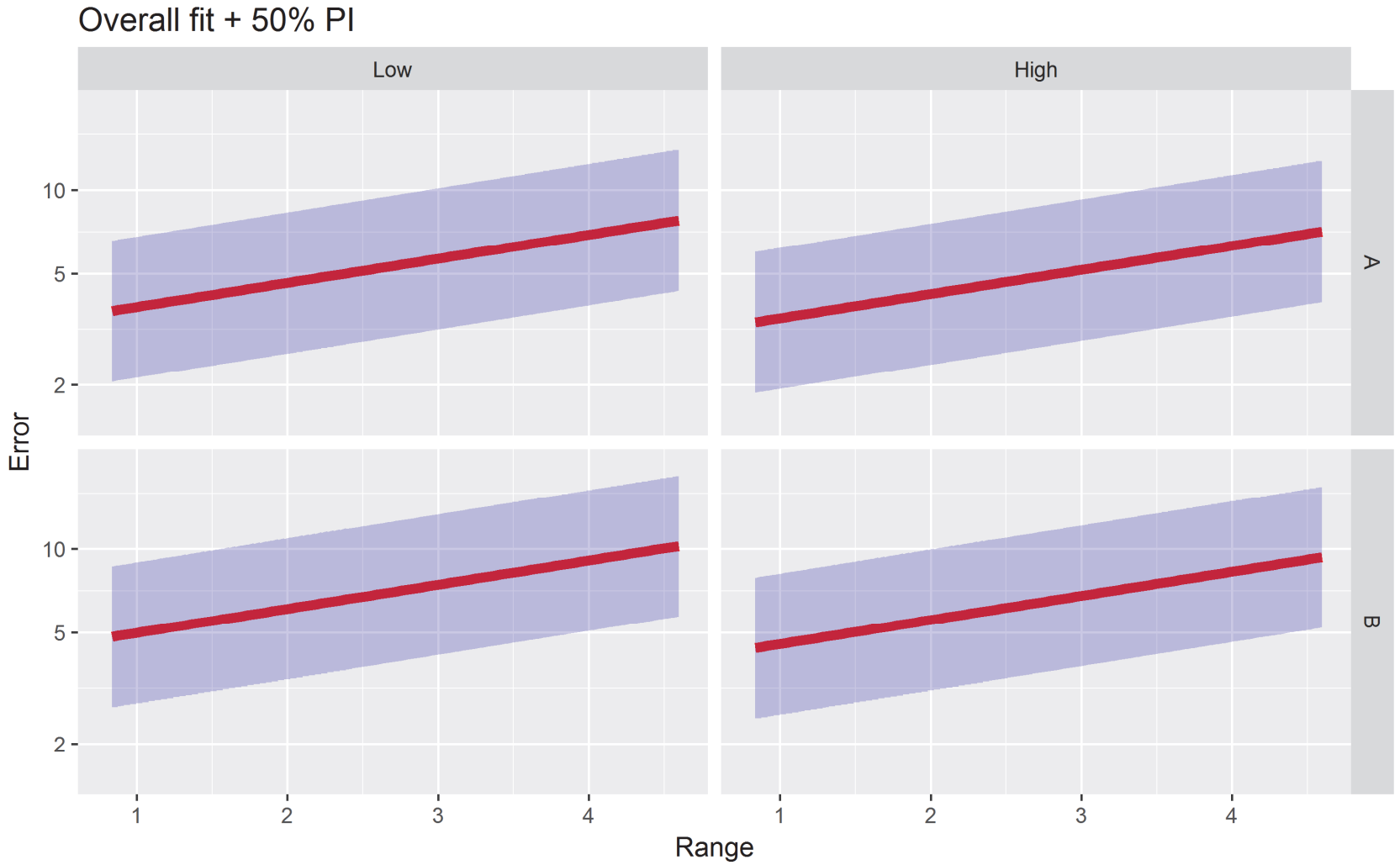
# Mixed model data analysis*

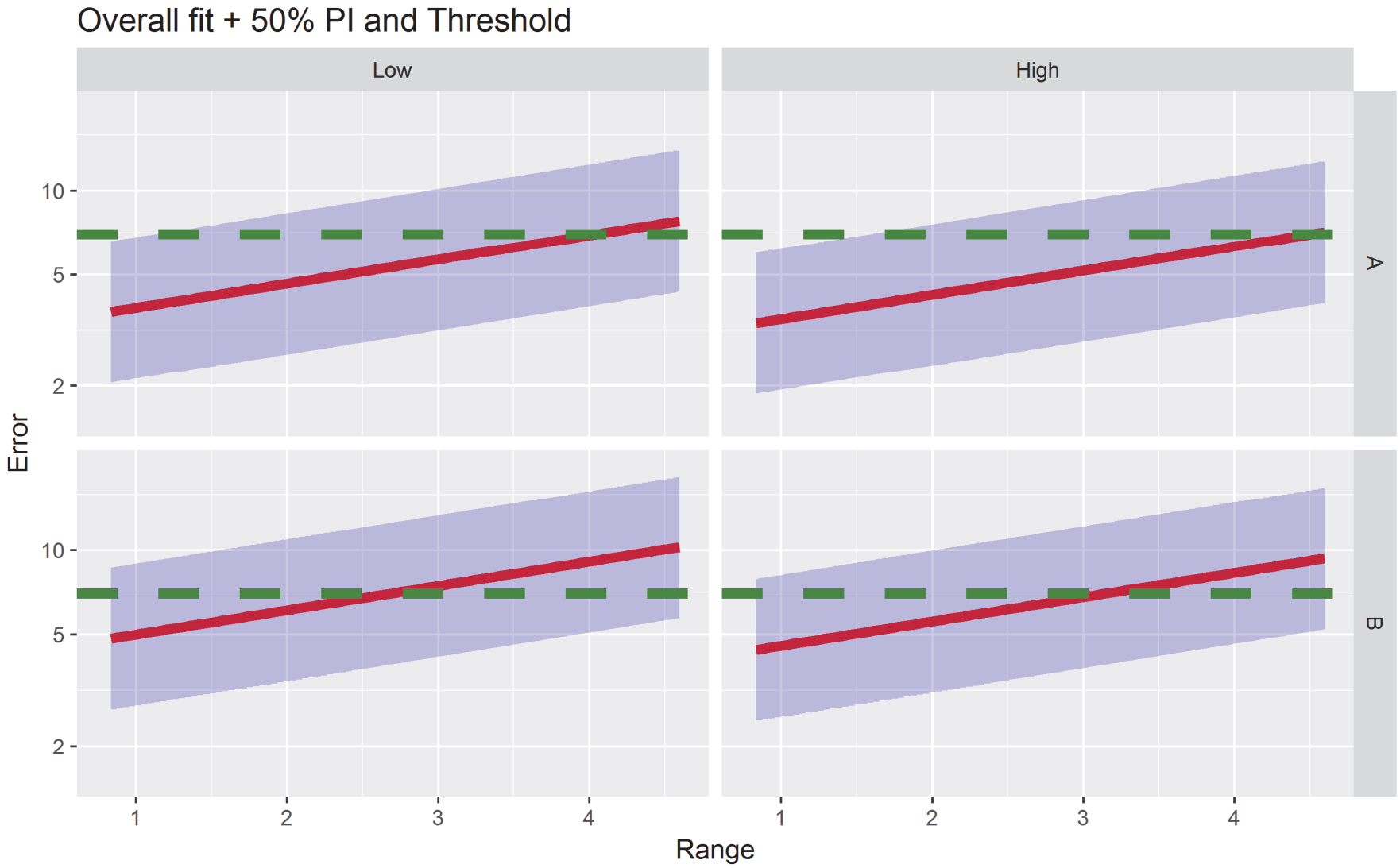
~100 lines of code, requires simulation

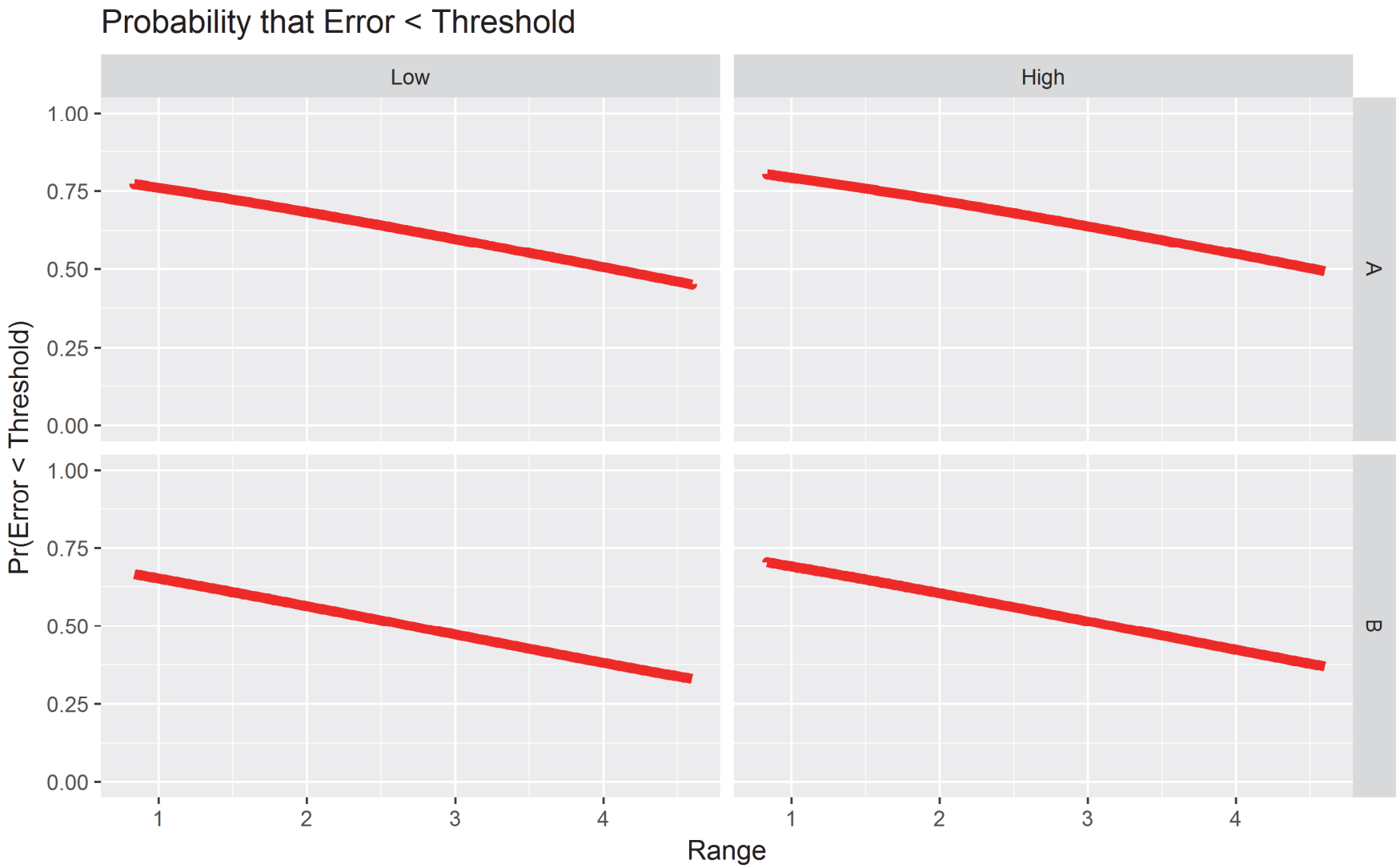



*True values and factor names obscured for data security.

~10 lines of code,
no simulation




Overall fit + 50% PI and Threshold
Low
High
A
B
Error
10
5
2
1
2
3
4
Range


Probability that Error < Threshold
Low
High
A
B
Pr(Error < Threshold)
1.00
0.75
0.50
0.25
0.00
1
2
3
4
Range

# The Scope of `ciTools`

# Current Status of `ciTools`

| | Confidence Intervals | Prediction Intervals | Probabilities | Quantiles |
|---|---|---|---|---|
| Linear Data | ✓ | ✓ | ✓ | ✓ |
| Count/ Binary Data | ✓ | ✓ | ✓ | ✓ |
| Skewed Data | ✓ | ✓ | ✓ | ✓ |
| Random Group Data | ✓ | ✓ | ✓ | ✓ |
| Reliability Data | Future Work | Future Work | Future Work | Future Work |
| Skewed, Random Group Data | In Progress | ✓ | ✓ | ✓ |
| … | … | … | … | … |

# Getting Help in `ciTools`

`?add_ci`

add_ci {ciTools} R Documentation

## Add Confidence Intervals for Predictions to Data Frames.

### Description

This is a generic function to append confidence intervals for predictions of a model fit to a data frame. A confidence interval is generated for the fitted value of each observation in `tb`. These confidence intervals are then appended to `tb` and returned to the user as a tibble.

### Usage

```
add_ci(tb, fit, alpha = 0.05, names = NULL, ...)
```

### Arguments

| | |
|---|---|
| `tb` | A tibble or data frame on which to make predictions. |
| `fit` | An object of class lm, glm, or lmerMod. Predictions are made with this object. |
| `alpha` | A real number between 0 and 1. Controls the confidence level of the interval estimates. |
| `names` | NULL or character vector of length two. If `NULL`, confidence bounds will automatically be named by `add_ci`, otherwise, the lower confidence bound will be named `ciNames[1]` and the upper confidence bound will be named `ciNames[2]`. |
| `...` | Additional arguments. |

### Details

For more specific information about the arguments that are useful in each method, consult

# Explore the math behind the scenes.

Uncertainty Intervals for Common Statistical Models

Literature Review

John Haman, Matt Avery

July 13, 2017

**Abstract**

The purpose of this document is to give an overview of all of the methods that are available in the package `ciTools`. In `ciTools`, one can supply our functions with a data set and a statistical model, and then `ciTools` will use either a default method or another method of similar quality to return the type of interval estimate they requested. We are not aware of another document that describes in detail the set of methods that one can use to produce interval estimates for each statistical model.

# `ciTools` in the larger R community

Open source, designed for collaboration and extension

1. Fits into the `tidyverse`.
2. Addresses existing gap in R capability with easy-to-learn syntax.
3. Use of generic functions makes growth easy.