Handling Uncertainty in Predictions

Approaches to Building Prediction Intervals Within a tidymodels Framework

Bryan Shalloway
Data Scientist, NetApp
2022-06-22

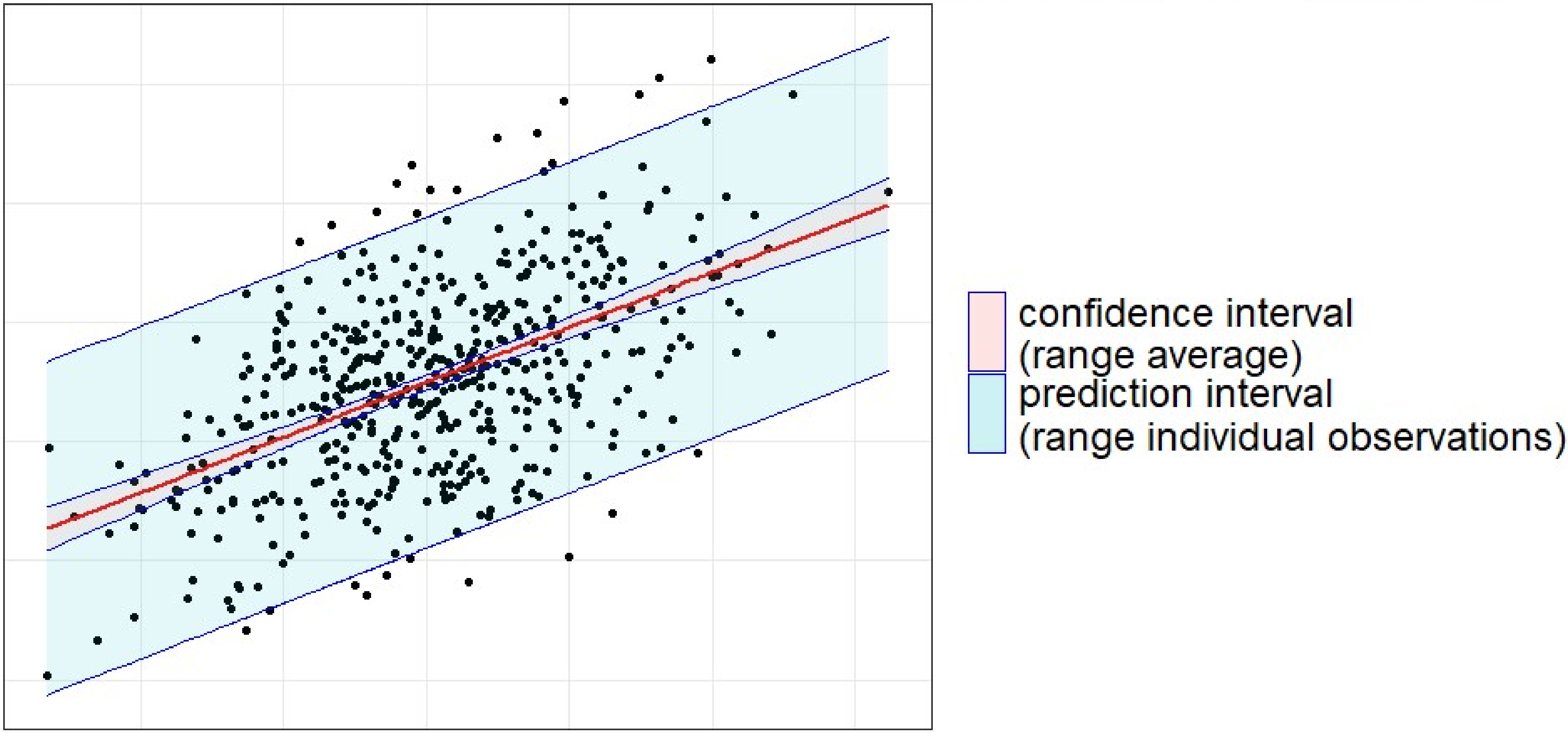
Write-ups:

- Part 1: Understanding Prediction Intervals
- Part 2: Simulating Prediction Intervals
- Part 3: Quantile Regression Forests for Prediction Intervals

tidymodels

@brshallo Bryan Shalloway bryanshalloway.com

Prediction Intervals are Wider Than Confidence Intervals



Model Building Steps (& Packages)

- Splitting Data
- Pre-processing
- Model specification
- Putting into a workflow
- Evaluating model

@brshallo Bryan Shalloway bryanshalloway.com

Linear Regression / Analytic Method

```r
bind_cols(
  predict(simple_wflow, test_data, type = "pred_int"),
  test_data
)
```

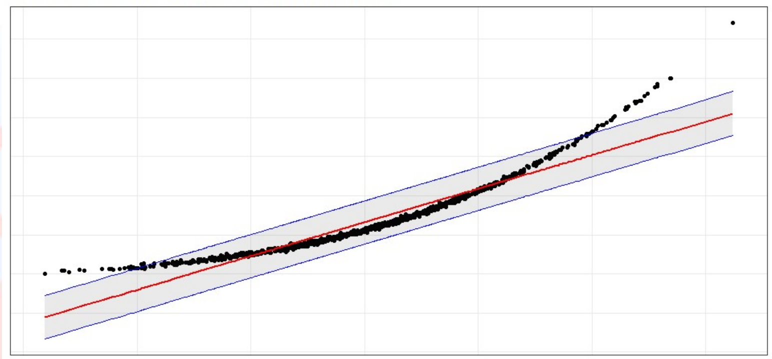
- Same thing for any modeling package interface that has a `type = "pred_int"` method. E.g. also works on Bayesian methods.

Limitations of Linear Regression

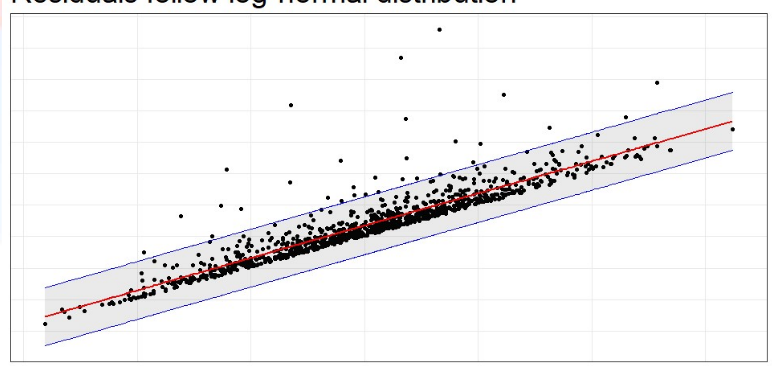
1. Many non-linear methods don't have a method available for prediction intervals.

2. Assumptions on distribution of residuals.

Model fit should be non-linear



Residuals follow log-normal distribution



Simulation based approach

```
library(workboots)

# 2000 bootstrap models
set.seed(345)
simple_pred_int <-
  simple_wf %>%
  predict_boots(
    n = 2000,
    training_data = train_data,
    new_data = test_data
  )
```

```
# summarise predictions with a 95% prediction interval
simple_pred_int %>%
  summarise_predictions()
#> # A tibble: 84 x 5
#>    rowid .preds               .pred .pred_lower .pred_upper
#>    <int> <list>               <dbl>       <dbl>       <dbl>
#>  1     1 <tibble [2,000 x 2]> 3465.       2913.       3994.
#>  2     2 <tibble [2,000 x 2]> 3535.       2982.       4100.
#>  3     3 <tibble [2,000 x 2]> 3604.       3050.       4187.
#>  4     4 <tibble [2,000 x 2]> 4157.       3477.       4764.
#>  5     5 <tibble [2,000 x 2]> 3868.       3305.       4372.
#>  6     6 <tibble [2,000 x 2]> 3519.       2996.       4078.
#>  7     7 <tibble [2,000 x 2]> 3435.       2914.       3954.
#>  8     8 <tibble [2,000 x 2]> 4072.       3483.       4653.
#>  9     9 <tibble [2,000 x 2]> 3445.       2926.       3966.
#> 10    10 <tibble [2,000 x 2]> 3405.       2876.       3938.
#> # ... with 74 more rowsturn go(f, seed, [])
}
```

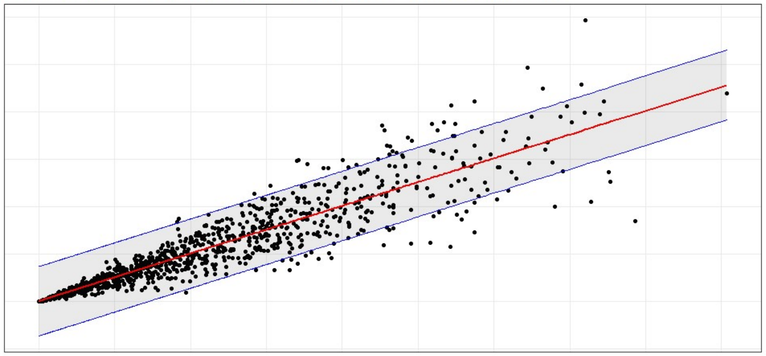
- Passing-in workflow, so allows pre-processing to influence widths.

** Also see conformal inference.*

Limitations of Simulation Based Approach

1. Consistent distribution of residuals

Assumes Consistent Variance of Residuals



2. Takes a long-time to run simulations

Quantile Regression

- Change objective function to optimize on quantiles
- E.g. predict 5th and 95th quantiles for a 90% prediction band
- Not available for all methods, but many popular approaches are available

RANGER

LightGBM

- Pass arguments in model specification.

```r
rf_mod <- rand_forest() %>%
  set_engine("ranger", ..., quantreg = TRUE) %>%
  set_mode("regression")
```

(Have to extract model from workflow for purposes of evaluation)

Write-ups

- Part 1: Understanding Prediction Intervals
  https://www.bryanshalloway.com/2021/03/18/intuition-on-uncertainty-of-predictions-introduction-to-prediction-intervals/
- Part 2: Simulating Prediction Intervals
  https://www.bryanshalloway.com/2021/04/05/simulating-prediction-intervals/
- Part 3: Quantile Regression Forests for Prediction Intervals
  https://www.bryanshalloway.com/2021/04/21/quantile-regression-forests-for-prediction-intervals/