Snow Gauge Calibration

Jiawei Yang, UC San Diego

1. Introduction

The Forest Service of United States Department of Agriculture uses a gamma transmission snow gauge to determine a depth profile of snow density, which helps with monitoring the water supply and flood management as well as studying climate change. However, the conversion functions of density values may change over the seasons.

The main goal is to provide a procedure to calibrate the snow gauge. To find an appropriate model, we carry out regression analysis on the raw data first. However, through graphical methods, we find a non-linear relationship between two variables, which implies data transformation. Based on the theoretical and empirical model, we do a log transformation and build a new model. We perform simulations to test the robustness of our model. We then perform forward prediction as well as reverse prediction by inverting the functions, where we produce point estimates and uncertainty bands for gains and densities. And we assess the accuracy of prediction by comparing the estimates and true values. In the cross-validation section, we run predictions by omitting two density values, respectively, and evaluate the prediction. In the advanced analysis, we carry out Ten-Fold Cross-Validation, which assesses the performance of Out-of-Sample R squared. Combining all evidence above, we conclude an appropriate conversion method.

Data:

The dataset is from a single calibration run of the snow gauge which consists of 90 observations: there are 10 measurements for each of 9 densities in grams per cubic centimeter of polyethylene. In this experiment, polyethylene blocks are used to simulate snow. The run measures the detected gamma photon count as “gain” between two poles of snow gauge of known densities.

2.1 Raw Data

Method:

To examine the necessity for data transformation, we first produce a scatter plot for raw data with density on the horizontal axis as explanatory variable and gain on the vertical axis as response variable. As for the linear regression analysis, we regress gain on density and present a table of coefficient statistics. We examine the reliability of this linear model by producing a residual plot and a normal q-q plot of residuals.

Analysis:

We produce a scatter plot with the regression line, a residual plot, and a q-q plot of residuals below. Also, we present the coefficient statistics of the linear regression model in the table below:

Figure I: Raw Data: Linear-Linear Regression, Residual Plot, and Q-Q plot of Residuals

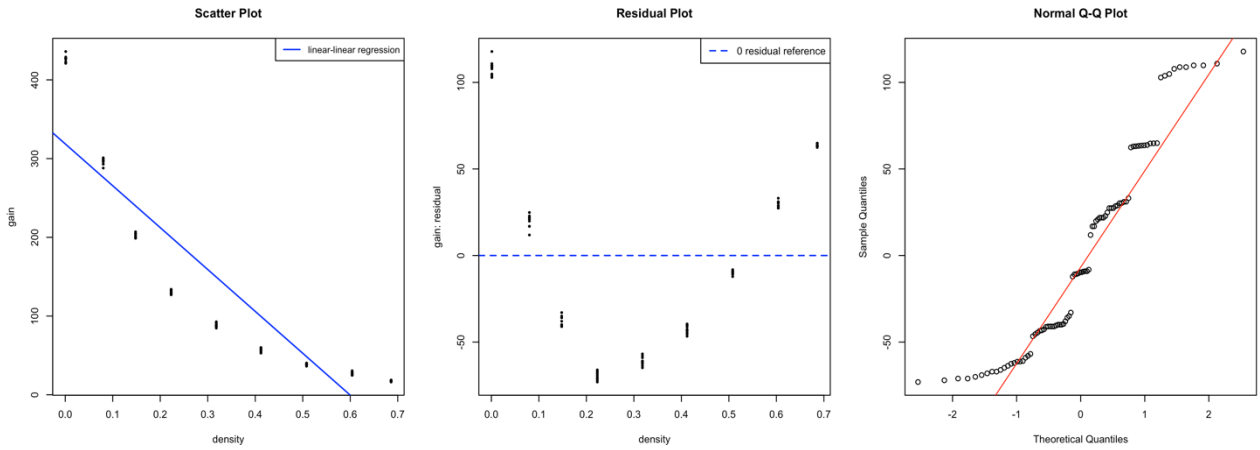


Table I : Coefficients of Regression (Raw Data)

	Intercept	Density	Adjusted R^2
Estimate	318.70*	-531.95*	0.8136
(Standard error)	(10.79)	(26.95)	

Note. * Significant at 1%

Despite the statistical significance of the estimated intercept and slope, as shown by the scatter plot, there is an obvious discrepancy between raw data trend and regression line where the regression line does not fit. Such an unfit regression line indicates a non-linear relationship between the two variables. Constant variability is violated as well, evidenced by the residual plot where residuals deviate from the zero lines. Besides, the residuals don't follow a normal distribution at all. The graphs suggest that it is inappropriate to apply linear regression models to raw data. Therefore, we need to come up with new models and transform the raw data.

Conclusion:

We conclude from the results that there is a non-linear relationship between density and gain, and it implies that transformation of data is needed to calibrate the snow gauge.

2.2 Transformed Data

Method:

To determine an appropriate data transformation method, we use the gamma ray measurement formula to identify the relationship between gain and density. Through log transformation of gain, we produce new scatterplot and residual plots using the transformed data. We carry out regression analysis on the new data and confirm our final model.

Analysis:

The physical model of gamma ray measurement provides intuition of log-transformation:

$$g = Ae^{\beta d}$$

We first take a log transformation on both sides:

$$\log(g) = \log\left(Ae^{\beta d}\right) = \log(A) + \log\left(e^{\beta d}\right)$$

Then we separate the terms according to product rule for exponentials. In the new equation, the explanatory variable is density d and the response variable becomes $\mathrm{Log}(g)$:

$$\mathrm{Log}(g) = \log(A) + \beta d$$

With the transformed data, we present the scatter plot with a regression line, the residual plot, and the q-q plot of residuals below. Besides, here is a table of coefficient statistics of the new regression model:

Figure II: Transformed Data: Linear-Linear Regression, Residual Plot, and Q-Q plot

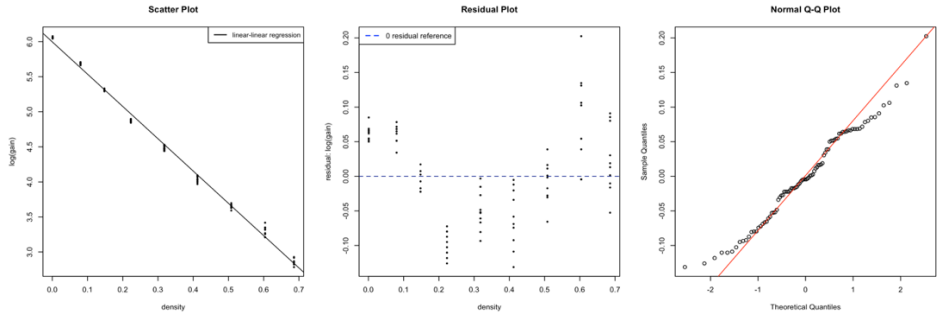


Table II : Regression Coefficients (Transformed Data)

	Intercept	Density	Adjusted R^2
Estimate	5.99727*	-4.60594	0.9958
(Standard error)	(0.01274)	(0.03182)	

Note. * Significant at 1%

While the coefficient statistics reveal a statistical significance of intercept and slope in this regression, the graphs in Figure 2 offer more insights about the appropriateness of our transformation: the scatterplot suggests a relatively linear relationship between the explanatory and the response variable; and the residual plot shows little deviations of residuals away the zero line, confirming a constant variability, and the q-q plot of residuals show a potential of normal distribution. Theoretically, our final method corresponds to the physical model of gamma-ray measurement, and through graphical analysis of the scatter plot, residual plot, and histogram, we gain empirical evidence that the model satisfies the conditions for the least-squares line.

Conclusion:

For our final method, we take a log transformation of gain according to the theoretical model. Through empirical results, we conclude that the model is suitable for linear regression analysis, given a fitter regression line and lower residual values.

2.3 Robustness against Measurement Error in Density: Simulation Results

Method:

We investigate the robustness of our slope estimate given the possibility of measurement errors in the density of the polyethylene blocks. We use the Monte Carlo simulation procedure to investigate the performance of estimates at different levels of error. Here we simulate random Gaussian errors.

$$\tilde{d} = d + \varepsilon \mid \varepsilon \sim N(0, s) \text{ for some } s \in \{0.001, 0.002 \cdots, 0.01\}$$

For each observation, we add a Gaussian error ε to the density, where the error follows a normal distribution centered at 0 with some pre-specified standard deviation *s*. We experiment with 10 values of *s*, ranging from 0.001 to 0.01, to represent the noise level. The error-appended data set is then denoted as $(\tilde{d}, log(g))$, where the response variables take values of log-transformed gains shown in section 2.2.

For each *s*, we repeat simulations to generate 1000 error-appended data sets. We investigate the same linear-log regression, and for the b^{th} data set we observe:

$$\widehat{\log(g)} = \hat{\beta}_1^{(b)} \tilde{d}_{(b)} + \hat{\beta}_0^{(b)}, \qquad b \in 1, 2, \cdots, 1000$$

These indicators are tracked to determine the overall performance of the fit across 1000 simulations at each noise level: percentage of slopes and intercept estimates significant at 5% and 10% level, and average R-square.

Analysis:

The expectation is a higher measurement error on average will introduce randomness in the model, which thus leads to a less-optimal fit than observed from the original data set. The simulation results are shown in the following Table III. (See visual demonstration in Appendix I)

Table III : Simulated Group Fit Performance Indicators (B = 1000)

	% significant ($\alpha = .05$)		% significant ($\alpha = .1$)		$\bar{R}^2$
Noise level *s* (10^{-3})	$\hat{\beta}_0^{(b)}$	$\hat{\beta}_1^{(b)}$	$\hat{\beta}_0^{(b)}$	$\hat{\beta}_1^{(b)}$	(%)
1	100	100	99.9	100	99.57

2	99.6	100	97.5	99.4	99.55
3	98.8	99.7	92.8	96.3	99.53
4	97.5	98.9	88.5	94.1	99.51
5	94.3	96.9	85.3	89.7	99.48
6	93.5	96.2	84.0	88.4	99.46
7	91.6	94.5	83.1	87.0	99.43
8	89.5	93.2	77.8	82.0	99.39
9	89.0	91.2	78.9	81.3	99.35
10	87.4	88.7	74.5	76.1	99.31

From Table III it is clear that all performance indicators are monotonically decreasing in the level of measurement error but at different rates. Evaluating the percentage of significant estimators, we claim that the slope estimate is more robust to measurement errors than the intercept estimates for both levels of the hypothesis testing: the significant percentages decrease at a slower rate. We also observe that the R-squared remains over 99% for all simulated groups with an increasing level of measurement error, which indicates a high predicting power of the model for changes in the gain in logarithmic scale responding to changes in density.

Conclusion:

According to the simulation result from Table I, we are able to conclude that the regression model is robust to Gaussian measurement errors with a standard deviation smaller than 0.003 by the evidence that over 90% of the slope and intercept estimates are significant at both levels of hypothesis testing. The predicting power, accessed by the R-squared index, remain high over 99% for all level of measurement error we simulate.

2.4 Forward Prediction

Method:

Our goal in this section is a forward prediction: ultimately, predicting the gain that corresponds to a certain measured density. The procedure is to first predict the gain in logarithmic scale with the following model we produce in section 2.2:

$$\hat{\log}(g) = 5.997 - 4.606d \qquad \hat{g} = e^{(5.997-4.606d)}$$

Since we are producing the interval estimate of a single observation of log(gain) given density, we use prediction interval of 95%, computed as:

$$\hat{\log}(g) \pm 2 \cdot sd(\log(g) - \hat{\log}(g))$$

The gain in original scale with error band is therefore obtained by:

$$e^{\hat{\log}(g) \pm 2 \cdot sd(\log(g) - \hat{\log}(g))}$$

Using transformed data, we plot a regression line and prediction bands on a scatter plot. We derive analytical equations of prediction bands by doing regression on the linear prediction band lines. We compare the point estimates and prediction intervals with the observed data and its range specifically for densities of 0.508 and 0.001 to determine the accuracy of prediction.

Analysis:

We produce a scatter plot with regression line and prediction bands below, and we also provide a table of point estimates and prediction bands for specific densities of 0.508 and 0.001.

Figure III: Point Estimates and Prediction Bands

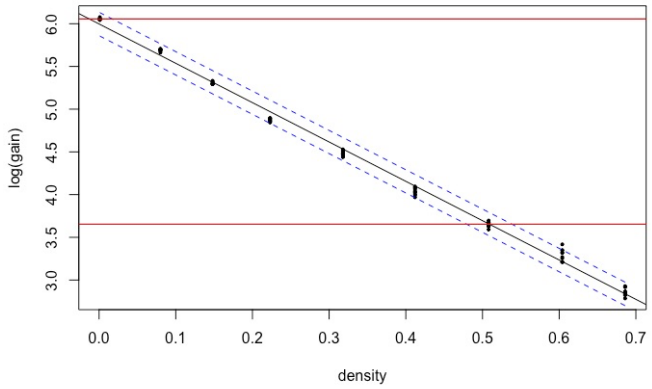


Table IV : Point Estimates and Uncertainty Bands (Forward)

	Point Estimate		Interval of Predicted Gain			Measured
Density	Log(Gain) (Band)	Gain	Lower	Upper	Interval Length	Range of Gain
0.001	5.993 (0.137)	400.479	349.094	459.425	55.1655	(421, 436)
0.508	3.657	38.762	33.827	44.417	5.295	(36.3, 40.3)

(0.136)

The analytic equation for prediction bands is acquired as followed:

$$\log(gain)_{lower} = 5.861 - 4.606d \quad gain_{lower} = e^{5.861-4.606d} = e^{-4.606d} * 351$$

$$\log(gain)_{upper} = 6.134 - 4.606d \quad gain_{upper} = e^{6.134-4.606d} = e^{-4.606d} * 461$$

According to the table, prediction bands for densities cover all points in the measured gain. Comparing the lengths of prediction bands, we observe that density with 0.508 would produce a narrower interval. Additionally, the range of measured gain shows that the group with density of 0.508 has a narrower range as well, which indicates higher accuracy. It suggests that some gains are predicted more accurately, in this case, which is the density group with 0.508.

Conclusion:

We conclude from the point estimates and uncertainty bands in the forward prediction that gains are predicted with different levels of accuracy. In this case, a density of 0.508 would produce a more accurate prediction band that covers all observed data in gain measurement.

2.5 Reverse Prediction

Method:

To achieve reverse prediction, we invert the forward prediction model produced in section 2.4, in which we use gain measurements to predict densities. We read the graph inversely and locate specific densities by gains. Using the reverse prediction model, we calculate the point estimates for densities that correspond to gain measurements of 38.6 and 426.7. As for the backward prediction bands, we invert the analytical equations above by reading the graph backward. We compare our results from reverse prediction with actual values and decide whether some densities are harder to be predicted.

Analysis:

Here is the forward prediction produced in section 2.4:

$$\hat{\log}(g) = 5.997 - 4.606d$$

By switching response and explanatory variables, we derive the reverse prediction equation:

$$\hat{d} = \frac{5.997 - \log(g)}{4.606}$$

$$\hat{d} = 1.302 - 0.217 \cdot \log(g)$$

According to Figure III, the prediction bands in reverse prediction would be the horizontal distance between the upper and lower band lines. We mark the corresponding prediction bands for gains of 38.6 and 426.7 as red in the figure. Here are the analytical equations of prediction bands:

$$d_{lower} = \frac{5.861 - \log(g)}{4.606} \qquad d_{upper} = \frac{6.134 - \log(g)}{4.606}$$

Using the reverse prediction model, we get the following estimates and prediction bands for densities that correspond to specific gain measurements of 38.6 and 426.7.

Table V : Point Estimates and Uncertainty Bands (Reverse)

		Actual Values	Point Estimate	Interval of Density	
Gain	Log(Gain)	Density	Density	Lower	Upper
38.6	3.653	0.508	0.509	0.479	0.538
426.7	6.056	0.001	-0.013	-0.042	0.017

By comparing the predicted and actual values, we find out that the point estimates are close to the true density values and true values do fall in the prediction intervals. However, from comparison, it shows that the point estimate from the gain of 38.6 is closer to its actual value, which indicates that some densities are harder to predict than others.

Conclusion:

Through reverse prediction, which we produce through inverting the forward prediction line, we conclude that the reverse prediction provides relatively accurate point estimates and prediction bands of densities. And it shows that some densities are harder to predict than others, given the fact that certain point estimates are closer to their observed values.

2.6 Cross-Validation

Method:

To avoid influences by certain measurements, we carry out cross-validations by two runs of

prediction which omits density values of 0.508 and 0.001, respectively. We apply the reverse prediction in section 2.5 to provide estimates and prediction intervals for density groups with average gain measurement of 38.6 and 426.7, which are densities of 0.508 and 0.001. We then compare the actual density with the predicted intervals.

Analysis:

Table VI below lists the point estimates and uncertainty bands produced from data omitting density values of 0.508 and 0.001, respectively.

Table VI : Point Estimates and Uncertainty Bands (Cross-Validation)

Gain Reading	Actual Density	Estimate Density	Interval of Density Lower	Interval of Density Upper
39.4	0.508	0.5047	0.4735	0.5360
37.6		0.5149	0.4836	0.5462
38.1		0.5120	0.4807	0.5433
37.7		0.5143	0.4830	0.5456
36.3		0.5225	0.4913	0.5538
38.7		0.5086	0.4773	0.5399
39.4		0.5047	0.4735	0.5360
38.8		0.5081	0.4768	0.5394
39.2		0.5058	0.4746	0.5371
40.3		0.4998	0.4685	0.5311
423	0.001	-0.0186	-0.0482	0.0110
421		-0.0176	-0.0472	0.0121
422		-0.0181	-0.0477	0.0115
428		-0.0212	-0.0508	0.0084
436		-0.0253	-0.0549	0.0043
427		-0.0207	-0.0503	0.0089
426		-0.0202	-0.0498	0.0094
428		-0.0212	-0.0508	0.0084
427		-0.0207	-0.0503	0.0089
429		-0.0217	-0.0513	0.0079

From the table, we observe that the actual density does fall in every prediction interval; for the group with average gain measurement of 38.6, the point estimates produce negative values in

density, and they are closer to the upper bound; Whereas in the other group with average of 426.7, there is no recognizable pattern of the locations within the intervals.

Conclusion:

Through running predictions omitting certain density block, we conclude that the actual density is covered in every prediction interval. And we discover that the actual density would be closer to upper band for groups with higher average gain measurement, namely the group with 38.6 in this comparison.

3. Advanced: Ten-Fold Cross Validation of Performance of Out-of-Sample R squared

Method:

The experiment is conducted with an equal number of observations at each level of density: in the original data set, 10 observations at 9 density levels with roughly equal distance between adjacent groups. Such pattern persists in our specific cross-validation in section 2.6 when observations are deleted by density. This specific grouping-by-density scattering may lead to deceptively high predicting power of linear models constructed thereby constructed. We therefore further generalize the cross-validation to a standard ten-fold case.

We exclude 20 observations subject to measurement errors mentioned in section 2.6, and therefore have validation groups of size 7 and training groups of 63. To investigate the behavior of the predictive power of the linear model at points not contained in the training set, we introduce out-of-sample R squared (R_{oos}) in two variations of Campbell and Thompson (2008):

$$R^2_{OOS(1)} = 1 - \frac{\sum_{i=1}^{7}(\log(g)_i - f_{train}(d_i))^2}{\sum_{i=1}^{7}(\log(g)_i - \overline{\log(g)}_{train})^2}$$

$$R^2_{OOS(2)} = 1 - \frac{\sum_{i=1}^{7}(\log(g)_i - f_{train}(d_i))^2}{\sum_{i=1}^{7}(\log(g)_i - f_{total}(d_i))^2}$$

Analysis:

R_{oos} ranges from minus infinity to 1 and is positively correlated with the out-of-sample predictive power of our model. We discover that the mean $R^2_{OOS(1)}$ and $R^2_{OOS(2)}$ are 99.47% and -5.11% across 10 sampled folds. They respectively indicate that the regression model from the

training set still explains 99.47% of variations in the log-gain for the unseen validation group, but still an observable less optimal performance than the original regression model constructed from all 70 data points.

Conclusion:

Proven by the ten-fold cross-validation, our log-linear model features consistently high out-of-sample predictability; However, it is still noticeable that the predictive power of the original regression model derived in section 2.2 may be exaggerated as a result of the organized scattering of the observations.

4. Summary and Discussion

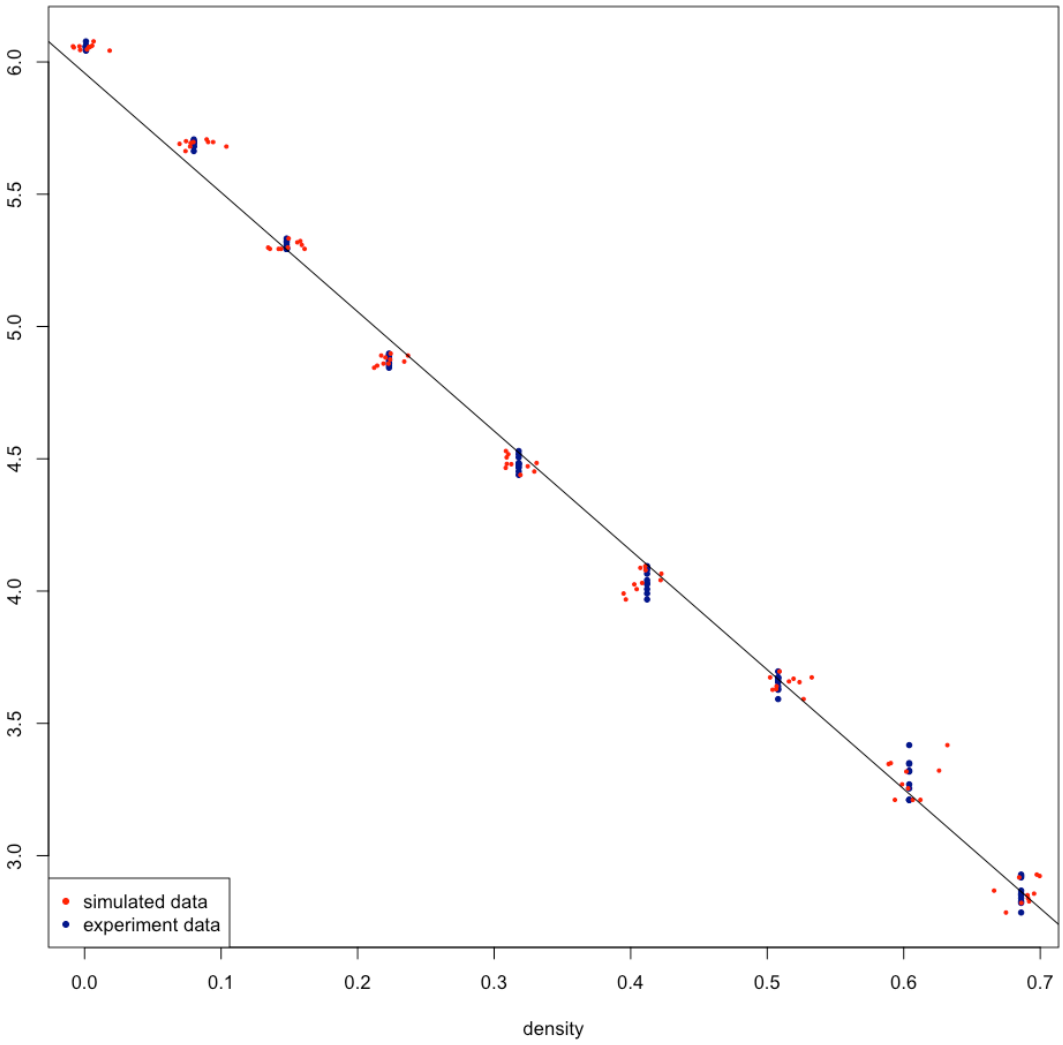
We have developed the calibrating equation for predicting snow density based on observed gamma-ray gain. This is done by inverting the regression model from the experiment where the gamma-ray gain is measured at a pre-determined density level of polyethylene. One should note the underlying physical theoretical model features an exponential relationship between the explanatory and response variables; We use logarithmic transformation on the gain variable to fit a linear model. We have tested the possible effect of measurement errors with Monte Carlo simulations, and out-of-sample predictive power with cross-validation: our regression model is shown to be robust again with some level of noise, and reliable in out-of-sample predicting.

We address the significance of the log-transformation: when it reduces calculation into linear form, potential drawbacks arise. Noticeably, observations with high gain values may produce less accurate predictions as taking the log of a large number reduces skewness to make it less identifiable from neighboring large numbers. Also, we notice that the prediction band relies on the condition of normal distribution of residuals, which may not be adequately satisfied in our model.

For further investigations, we would suggest experimenting with an increased level of accuracy in raw data measurement for the density level. Also, it could be potentially beneficial to reduce the clustering of the observations by taking more possible density values to avoid close scattering of leverage or influential points in case of systematic experiment or measurement error.

5. Appendices

Appendix I: Sample Measurement Error-Appended data (error~N(0,0.01))



6. Citation

Campbell, John Y., and Samuel B. Thompson. “Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average?” The Review of Financial Studies, vol. 21, no. 4, 2008, pp. 1509–1531. JSTOR, www.jstor.org/stable/40056860. Accessed 24 May 2021.