Snow

Kaisa Roggeveen, Scott Graham

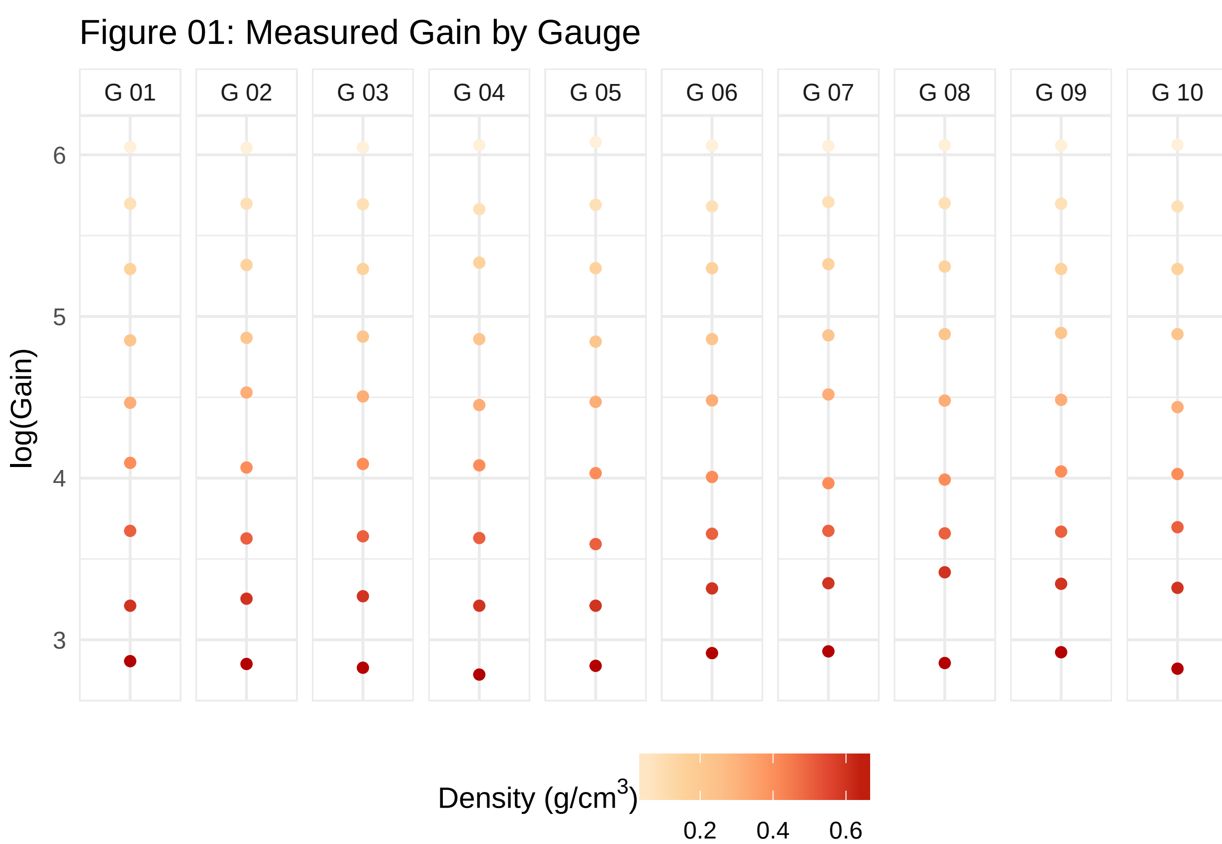
April 6th, 2018

Introduction

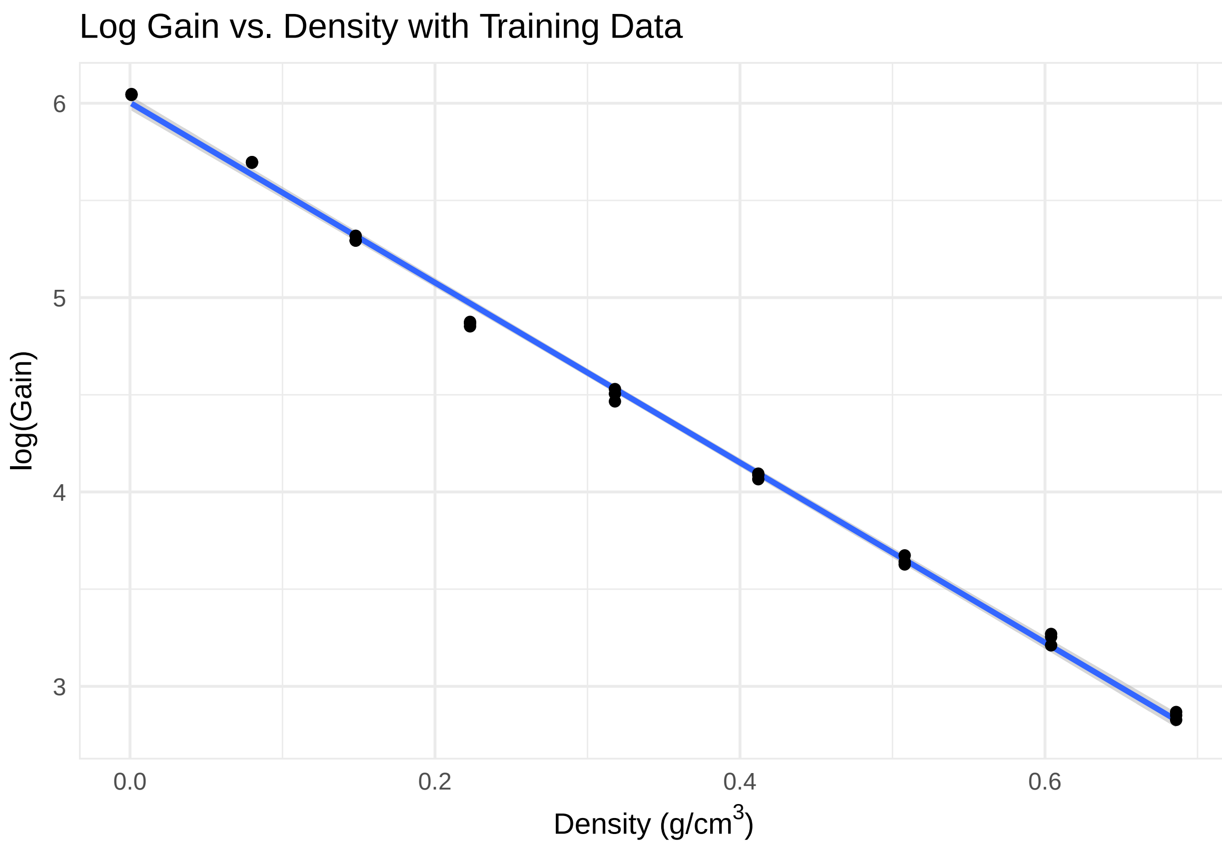
Data

Table 1: Wide Data

Density	G 01	G 02	G 03	G 04	G 05	G 06	G 07	G 08	G 09	G 10
0.686	17.6	17.3	16.9	16.2	17.1	18.5	18.7	17.4	18.6	16.8
0.604	24.8	25.9	26.3	24.8	24.8	27.6	28.5	30.5	28.4	27.7
0.508	39.4	37.6	38.1	37.7	36.3	38.7	39.4	38.8	39.2	40.3
0.412	60.0	58.3	59.6	59.1	56.3	55.0	52.9	54.1	56.9	56.0
0.318	87.0	92.7	90.5	85.8	87.5	88.3	91.6	88.2	88.6	84.7
0.223	128.0	130.0	131.0	129.0	127.0	129.0	132.0	133.0	134.0	133.0
0.148	199.0	204.0	199.0	207.0	200.0	200.0	205.0	202.0	199.0	199.0
0.080	298.0	298.0	297.0	288.0	296.0	293.0	301.0	299.0	298.0	293.0
0.001	423.0	421.0	422.0	428.0	436.0	427.0	426.0	428.0	427.0	429.0



Talk about how the gauges are consistent



talk about how a log model is appropriate

Training vs. Validation Data

Talk about how we split the data into training data (Gauges 1-3), and validation data (gauges 4-10).

Calibration

Classic Calibration Method

For the classic calibration method, we regress our measurement (G := Gain) as a function of the known variable (D := Density).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.003	0.01799	333.6	3.897e-47
Density	-4.63	0.04495	-103	2.182e-34

Table 3: Fitting linear model: log(Gain) ~ Density

Observations	Residual Std. Error	R^2	Adjusted R^2
27	0.05255	0.9976	0.9976

From this we take our linear regression model, and invert it, solving for the known predictor variable D. This gives us:

$$\hat{D}_i = -\frac{\ln(G_i) - 6.0032 - \epsilon_i}{4.6301} = 1.2965(1 + \epsilon_i) - 0.2160 \ln(G_i), \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

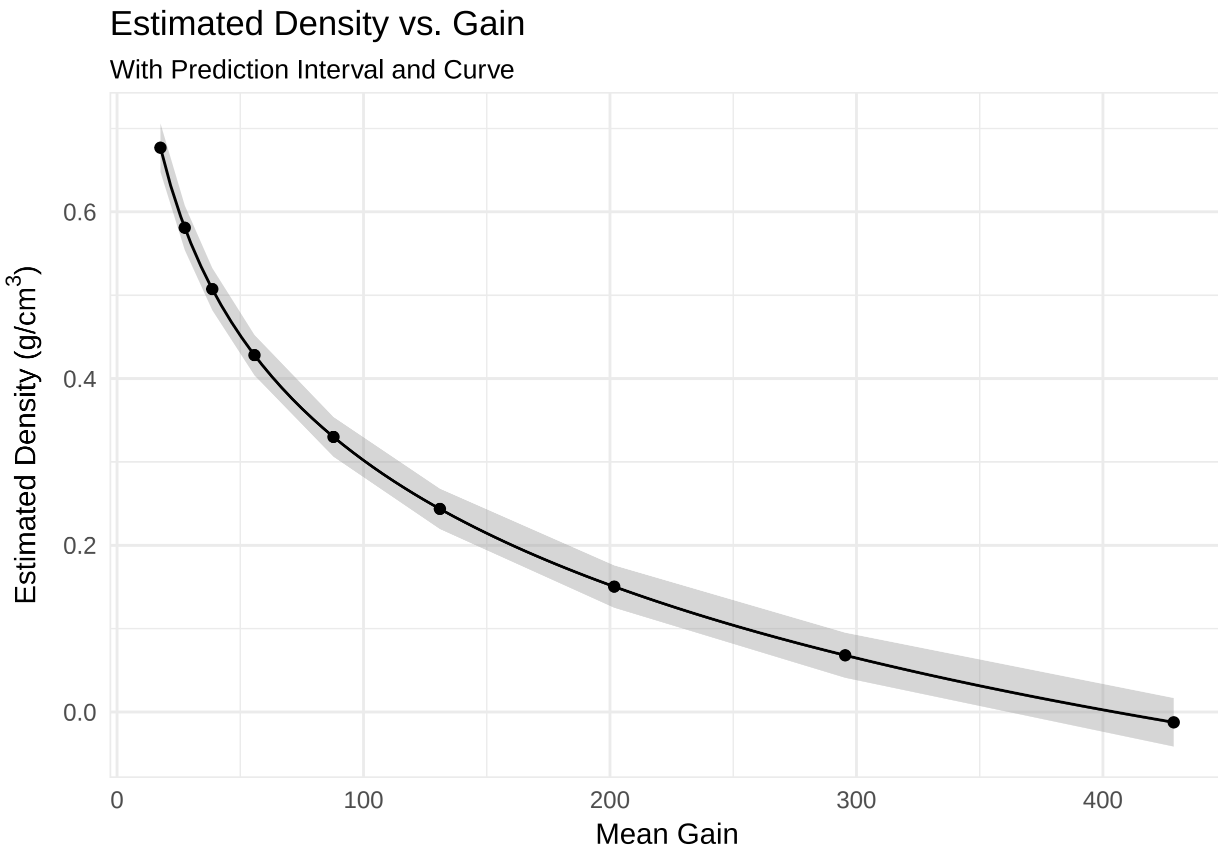
From this, we can come up with both the point estimates for D, and a prediction interval, using:

$$\text{se}\left(\hat{D}_i\right) = \frac{\sqrt{MSE}}{\hat{\beta}_1}\sqrt{1 + \frac{1}{n} + \frac{\left(D_i - \bar{D}\right)^2}{S_{DD}}}, MSE = 0.0028, \bar{D} = 0.3311\ldots, S_{DD} = 0.2293$$

As well as assuming an underlying Student's t-distribution, with $df = n - p = n - 2 = 27 - 2 = 25$.

Density	Mean Gain	Est Density	Prediction Std. Error	Prediction LB	Prediction UB
0.001	428.71429	-0.0124462	0.0141394	-0.0415669	0.0166745
0.080	295.42857	0.0679760	0.0131343	0.0409254	0.0950265
0.148	201.71429	0.1503875	0.0123271	0.1249995	0.1757756
0.223	131.00000	0.2436153	0.0117434	0.2194293	0.2678012
0.318	87.81429	0.3300005	0.0115588	0.3061946	0.3538063
0.412	55.75714	0.4281015	0.0117852	0.4038294	0.4523736
0.508	38.62857	0.5073681	0.0122907	0.4820549	0.5326812
0.604	27.47143	0.5809831	0.0129879	0.5542339	0.6077322
0.686	17.61429	0.6769713	0.0141709	0.6477857	0.7061569

From this, we can plot the results, with the Estimated Density on the y-axis, and the Mean Gain for a given Density on the x-axis. The grey band represents the 95% Prediction Interval for D.

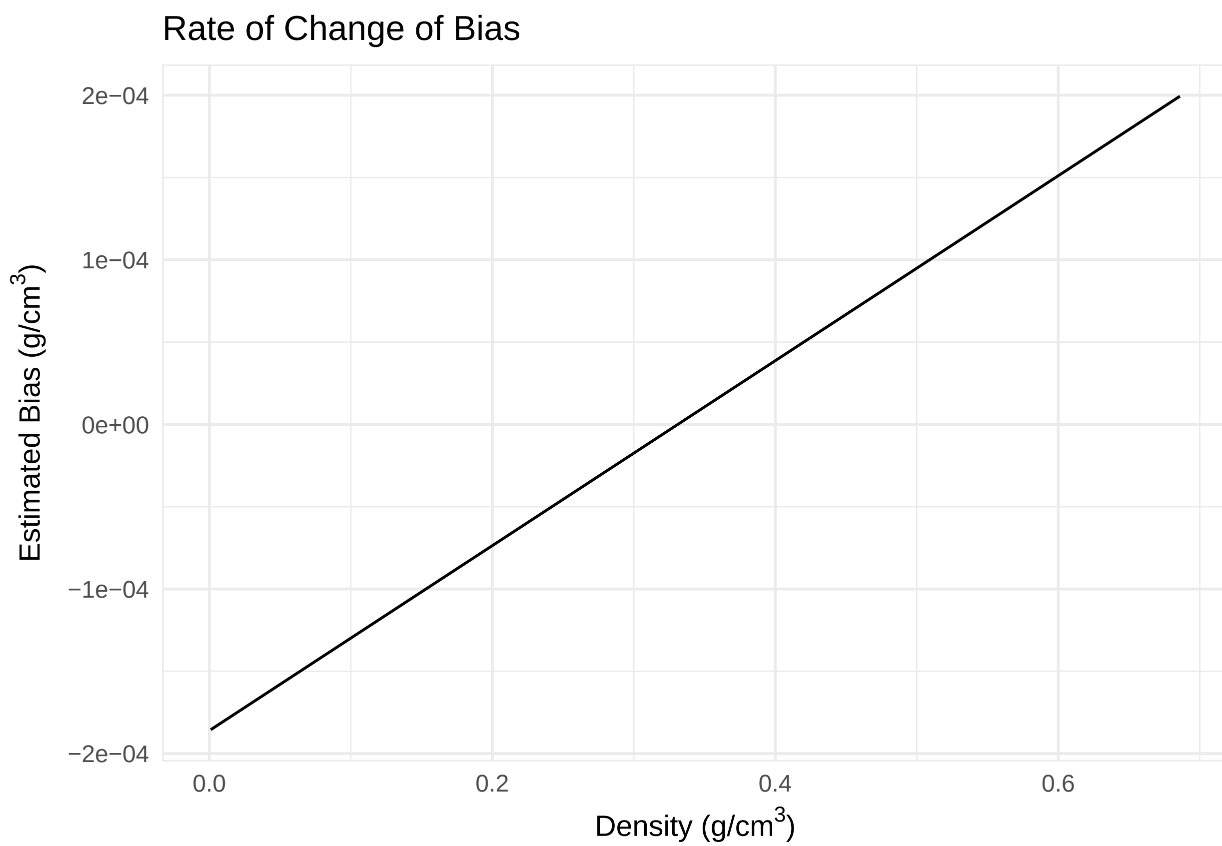


Parker et al. (2010), provides a method for finding a finding the bias in the estimation of $\hat{D}_i$:

$$\text{bias}\left(\hat{D}_i\right) = \frac{(D_i - \bar{D})MSE}{\hat{\beta}_1^{\,2} S_{DD}}$$

Table 5: Unbiasing the Estimated Density for Classic Calibration

Density	Est Density	Bias	Unbiased Est Density
0.001	-0.0124462	-0.0001855	-0.0122607
0.080	0.0679760	-0.0001411	0.0681171
0.148	0.1503875	-0.0001029	0.1504904
0.223	0.2436153	-0.0000607	0.2436760
0.318	0.3300005	-0.0000074	0.3300078
0.412	0.4281015	0.0000454	0.4280560
0.508	0.5073681	0.0000994	0.5072687
0.604	0.5809831	0.0001533	0.5808297
0.686	0.6769713	0.0001994	0.6767719



Inverse Regression

The other methodology looked at was using a inverse regression technique, by using D as the dependent variable, and G as the independent variable. This should give us a slightly different result than the coefficients calculated under the classical calibration method, as regression equations don't invert exactly unless the correlation between the two variables is ± 1. Because our correlation (-0.9061202) is close to -1, the coefficients will be very close. For this we use the model:

$$\hat{D}_i = \hat{\gamma_0} + \hat{\gamma_1}\left(\ln(G_i) - \overline{\ln(G_i)}\right) + \epsilon_i = 0.3311 - 02155\left(\ln(G_i) - 4.4701\right) + \epsilon_i, \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\overline{\ln(G)} = \sum_{i=1}^{n} \frac{\ln G_i}{n}, \hat{\gamma_0} = \bar{D}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3311	0.002182	151.8	1.375e-38
log(`Centred Gain`)	-0.2155	0.002092	-103	2.182e-34

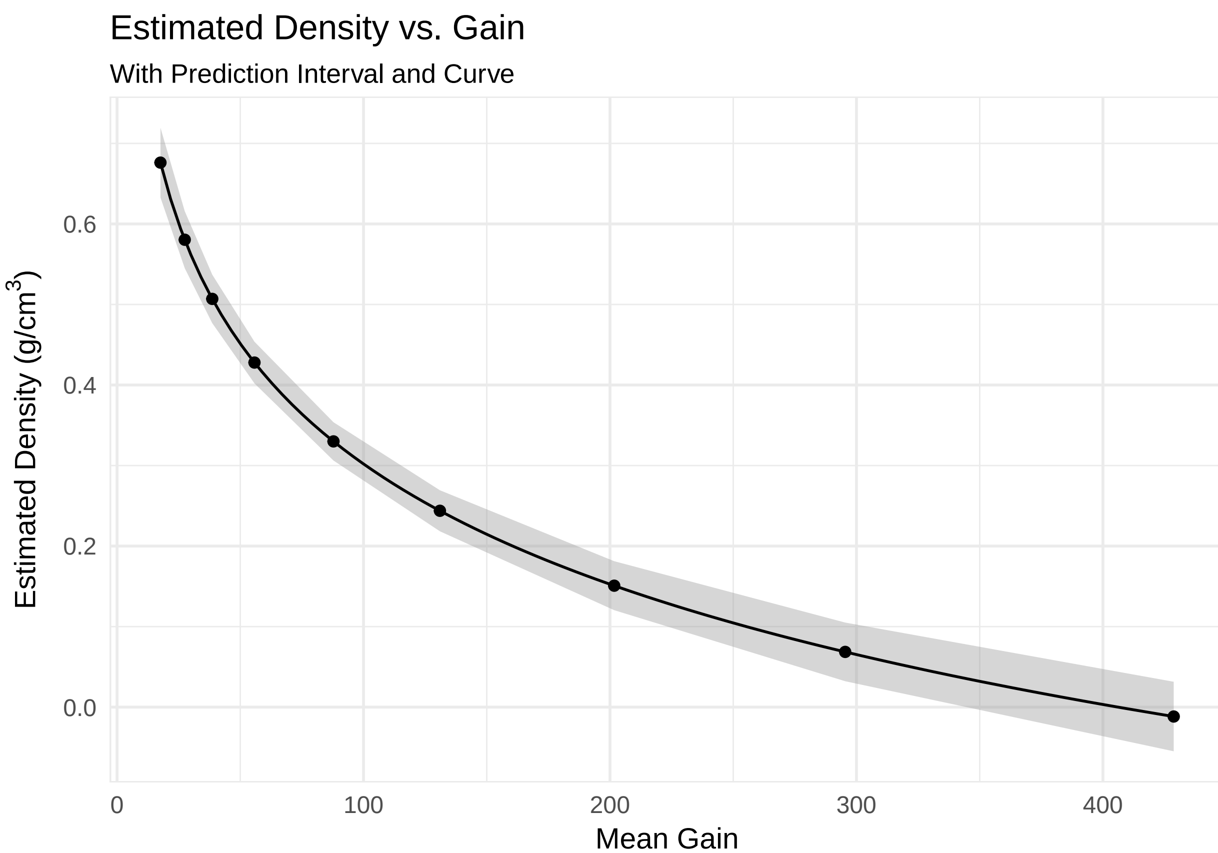
Table 7: Fitting linear model: Density ~ log(`Centred Gain`)

Observations	Residual Std. Error	R^2	Adjusted R^2
27	0.01134	0.9976	0.9976

As we can see from the regression output, $\hat{\gamma}_0 = \bar{D}$, which is what we want. We then estimated $\hat{D}_i$, and the prediction interval estimate:

Density	Mean Gain	Mean Centred Gain	Est Density	Prediction Std. Error	Prediction LB	Prediction UB
0.001	428.71429	4.9072264	-0.0116386	0.0209588	-0.0548040	0.0315268
0.080	295.42857	3.3815876	0.0685945	0.0176857	0.0321701	0.1050190
0.148	201.71429	2.3088983	0.1508124	0.0147635	0.1204064	0.1812183
0.223	131.00000	1.4994757	0.2438210	0.0123749	0.2183344	0.2693075
0.318	87.81429	1.0051556	0.3300031	0.0115453	0.3062250	0.3537811
0.412	55.75714	0.6382174	0.4278735	0.0125570	0.4020119	0.4537351
0.508	38.62857	0.4421573	0.5069538	0.0146228	0.4768375	0.5370700
0.604	27.47143	0.3144484	0.5803957	0.0171798	0.5450132	0.6157782
0.686	17.61429	0.2016198	0.6761583	0.0210567	0.6327912	0.7195254

As well as recreating the plot used in the classical calibration method. Note that the plot looks very similar, with a marginally large standard error term.



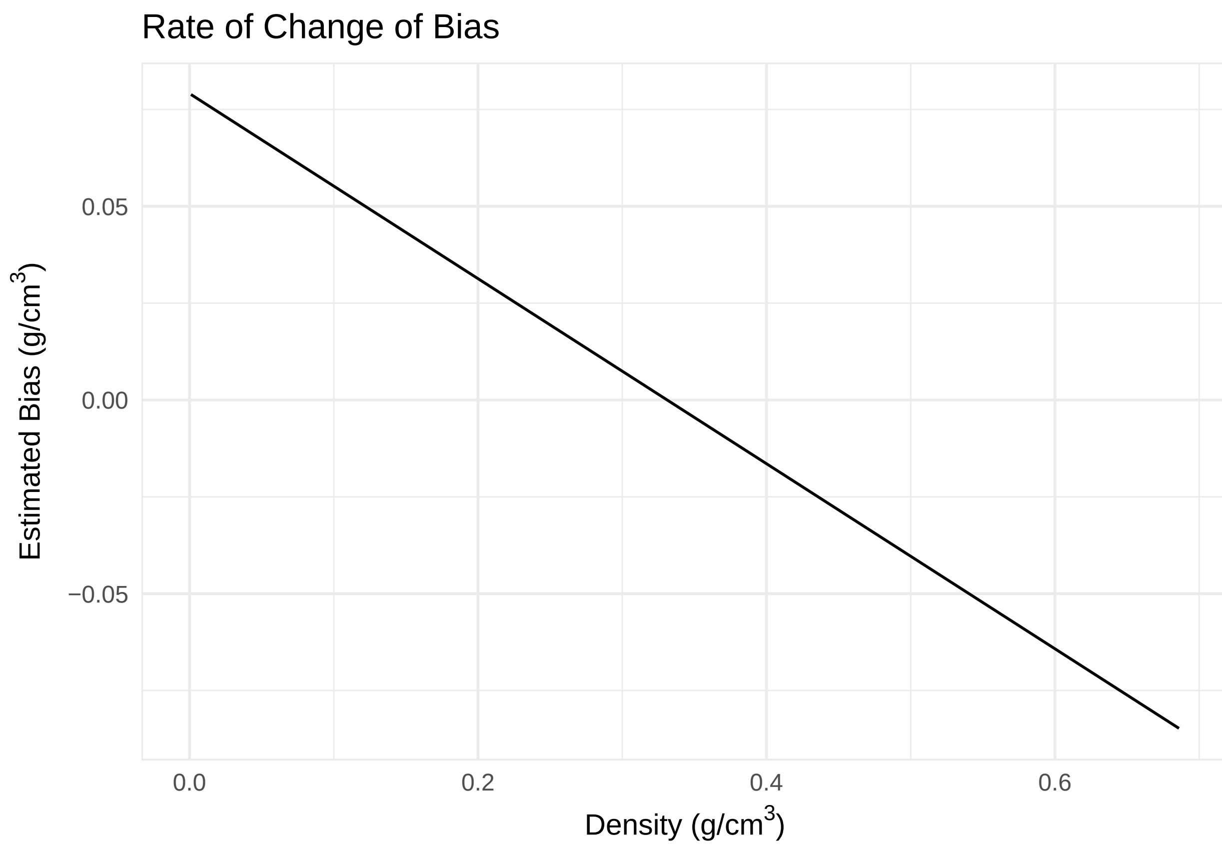
Parker et al. (2010), provides a means of unbiasing $\hat{D}_i$ for inverse regression too:

$$\text{bias}\left(\hat{D}_i\right) = \frac{\bar{D} - D_i}{1 + \frac{\hat{\beta_1}^2 S_{DD}}{(n-1)MSE}}$$

This is then plotted against D, to see the rate of change.

Table 9: Unbiasing the Estimated Density for Inverse Regression

Density	Est Density	Bias	Unbiased Est Density
0.001	-0.0116386	0.0788687	-0.0905073
0.080	0.0685945	0.0599943	0.0086002
0.148	0.1508124	0.0437481	0.1070643
0.223	0.2438210	0.0258294	0.2179915
0.318	0.3300031	0.0031324	0.3268706
0.412	0.4278735	-0.0193256	0.4471991
0.508	0.5069538	-0.0422615	0.5492153
0.604	0.5803957	-0.0651974	0.6455931
0.686	0.6761583	-0.0847885	0.7609467



Comparison

Table 10: Comparison of Bias

Density	CC Est Density	CC Bias	IR Est Density	IR Bias
0.001	-0.0124462	-0.0001855	-0.0116386	0.0788687
0.080	0.0679760	-0.0001411	0.0685945	0.0599943
0.148	0.1503875	-0.0001029	0.1508124	0.0437481
0.223	0.2436153	-0.0000607	0.2438210	0.0258294
0.318	0.3300005	-0.0000074	0.3300031	0.0031324

Density	CC Est Density	CC Bias	IR Est Density	IR Bias
0.412	0.4281015	0.0000454	0.4278735	-0.0193256
0.508	0.5073681	0.0000994	0.5069538	-0.0422615
0.604	0.5809831	0.0001533	0.5803957	-0.0651974
0.686	0.6769713	0.0001994	0.6761583	-0.0847885

Table 11: Comparison of Standard Error

Density	Mean Gain	CC Est Density	CC Prediction Std. Error	IR Est Density	IR Prediction Std. Error
0.001	428.71429	-0.0124462	0.0141394	-0.0116386	0.0209588
0.080	295.42857	0.0679760	0.0131343	0.0685945	0.0176857
0.148	201.71429	0.1503875	0.0123271	0.1508124	0.0147635
0.223	131.00000	0.2436153	0.0117434	0.2438210	0.0123749
0.318	87.81429	0.3300005	0.0115588	0.3300031	0.0115453
0.412	55.75714	0.4281015	0.0117852	0.4278735	0.0125570
0.508	38.62857	0.5073681	0.0122907	0.5069538	0.0146228
0.604	27.47143	0.5809831	0.0129879	0.5803957	0.0171798
0.686	17.61429	0.6769713	0.0141709	0.6761583	0.0210567

To compare the two methods, we looked at the size of their Bias, and the size of the Standard Errors. From the two tables above, one can see that the Classic Calibration method outperformed the Inverse Regression by having both a smaller estimated bias, and a smaller estimated standard error based on the same training and validation samples.

Measurement Error

If we were to assume that the given densities for the polyethylene blocks contained small amounts of measurement error, this change the size of our interval estimates.

Let:

$$\hat{D}_i = D_i + \epsilon_{D,i}, \epsilon_{D,i} \sim \mathcal{N}(0, \sigma^2) \implies$$

$$\ln(G_i) = \hat{\beta}_0 + \hat{\beta}_1 \hat{D}_i + \epsilon_{G,i} = \hat{\beta}_0 + \hat{\beta}_1 (D_i + \epsilon_{D,i}) + \epsilon_{G,i} =$$

$$\hat{\beta}_0 + \hat{\beta}_1 D_i + (\hat{\beta}_1 \epsilon_{D,i} + \epsilon_{G,i}) = \hat{\beta}_0 + \hat{\beta}_1 D_i + \epsilon^{\star}_{G,i}$$

Where

$$\epsilon^{\star}_{G,i} \sim \mathcal{N}\left(0, \hat{\beta}_1^{\,2} \sigma_D^2 + \sigma_G^2 + 2\hat{\beta}_1 \operatorname{Cov}(D, G)\right)$$

Now hopefully the covariance term is equal to 0, otherwise additional issues would arise in the calibration. While this won't affect the coefficient estimation done in the regressions, it would affect the size of the interval estimates, by increasing them to reflect the greater uncertainty in the quality of measurements.