Contributions to Variation in Fly Ball Distances: A Followup

Alan Nathan*

University of Illinois at Urbana-Champaign

(Dated: July 20, 2020)

I. INTRODUCTION

A couple of weeks ago, my article Contributions to Variation in Fly Ball Distance appeared in FanGraphs. The purpose of the article was to answer a question I had long ago posed about the factors that result in the variation of fly ball distances. The closing sentence of that paper was, "It is now time to put this baby to rest." But I didn't put it to rest. I received a number of interesting private communications offering suggestions for further analysis. So, here I am again!

II. A NEW ANALYSIS

Following both the suggestions as well as some new thoughts of my own, I decided to redo the analysis with some significant modifications.

1. The data set was enlarged considerably, from 719 to 4697 batted balls, a 6.5-fold increase. This was achieved in part by including data not only from Tropicana Field but also from six additional stadiums with a retractable roof, but only for games in which the roof was closed: Chase Field, Marlins Park, Miller Park, Minute Maid Park, Rogers Centre, and T-Mobile Park. In addition, it included data not only from the 2019 season but from 2016-2018 as well. Finally, the range of launch angles θ was expanded from 25°-30° to 15°-40°, keeping the range of exit velocities v_0=94-112 mph. This change allows investigation of batted balls of varied character. As before, only batted balls were included in which the tracked distance was at least 80% of the total distance.

2. A split-data analysis was done, with half of the data randomly selected as "training

* email: a-nathan@illinois.edu; Twitter: @pobguy; Web: baseball.physics.illinois.edu

data" and the remaining half used as "test data". The models were established by fitting to the training data and evaluated based on how well they described the test data. This analysis is discussed in Sec. III.

3. The physical variables that constituted the model were modified as follows:

   - Since the batted ball distance is expected to vary greatly over the increased range of θ, that variable needs to be part of the model.
   - While the exit velocity v_0, the backspin ω_b, the sidespin ω_s, and the launch angle θ may depend on the adjusted spray angle ϕ_1,[1] there is no reason why a batted ball distance should otherwise depend on it.[2] Therefore, ϕ_1 was eliminated as a model parameter.
   - The is no physical reason why the batted ball distance should depend on the *sign* of ω_s. Therefore ω_s was eliminated as a model parameter and replaced by the absolute value $\mid \omega_s \mid$.

4. In the earlier study some fraction of the variation in batted-ball distances was attributed to noise. I take a closer look at the factors contributing to the noise in Sec.IV.

5. As before the spin-independent drag coefficient C_{D0} was determined from analysis of the first $\sim$2 sec of batted-ball trajectory. As will be discussed in Sec. V, a separate study was done to compare these values with those obtained from the corresponding pitched-ball trajectory. Also presented is a discussion of why it is better to use the batted-ball value.

III. THE SPLIT-DATA ANALYSIS

The results of the random split-data analysis is shown in Table I, in which the outcome of the sequence of models is presented. Each of Models 1-4 is a non-parameteric generalized

[1] Recall that the adjusted spray angle is such that it is negative for balls hit to the pull field and positive for balls hit to the opposite field, independent of the handedness of the batter.

[2] While wind could introduce a dependence on ϕ_1, I am only considering data from closed stadiums.

additive model (GAM). Model 5 will be discussed in Sec. IV. Some comments on the GAM models follow:

- There is virtually no difference in the R^2 or residual rms values for the training and test data, giving us confidence that the the GAM's have predictive value and that there is no "overfitting".
- Comparing Model 1 to Model 2, one sees that including the total spin ω produces very little net improvement to the fit.
- On the other hand, breaking ω into its constituent parts (Model 3) results in a significant improvement, reducing the residuals by more than a factor of two. The lesson learned is simple: Knowing the total spin rate is not enough; one also needs to know the spin axis. This conclusion was not obvious in the previous analysis due to the inclusion of ϕ_1 as a fitting parameter.
- Finally, including C_{D0} (Model 4) produces only a marginal improvement, suggesting that variation of the drag plays only a small role in comparison to the residual random measurement noise. I postpone further discussion of this feature until Sec. IV.

TABLE I: Model fits to batted-ball distance, where R^2 is the square of the Pearson correlation coefficient and rms is the root-mean-square deviation of the fit from the data. The fits were done on the training data, then applied to the test data. Models 1-4 utilize a GAM, whereas Model 5 is a Deming regression applied to the residuals of Model 3.

		Training		Test	
Model	Parameters	R^2	rms (ft)	R^2	rms (ft)
1	v_0+θ	0.678	23.9	0.682	23.9
2	v_0+θ+ω	0.706	22.8	0.706	23.1
3	v_0+θ+ω_b+$\mid \omega_s \mid$	0.935	10.6	0.937	10.6
4	v_0+θ+ω_b+$\mid \omega_s \mid$+C_{D0}	0.953	9.1	0.955	9.1
5	Model 3 + Deming on C_{D0}	—	—	0.951	9.5

Plots comparing fitted to actual distances for Models 3 and 4 are shown in Fig. 1, clearly showing the role played by C_{D0}. A plot of the Model 4 residuals is given in Fig. 2, which indicates that the model accurately describes the data to ± 2 ft over the range of distances

320-420 ft. Finally, Fig. 3 shows the Model 4 fits of batted-ball distances as a function of various parameters, with other parameters fixed. Some interesting features of Fig. 3 are worth pointing out:

- Both the launch angle and the backspin rate that maximize the distance decrease with increasing exit velocity. In both cases, this behavior is the result of the increase in drag with velocity.

- With other parameters fixed, the distance decreases with increasing C_{D0}, as one intuitively expects. Interestingly, the rate of decrease is greater at higher exit velocity. Once again, this behavior is the results of the increase in drag with exit velocity.

- The dependence of distance on exit velocity is approximately linear.

- For reasons discussed in an earlier article, batted-ball distance decreases with increasing sidespin.

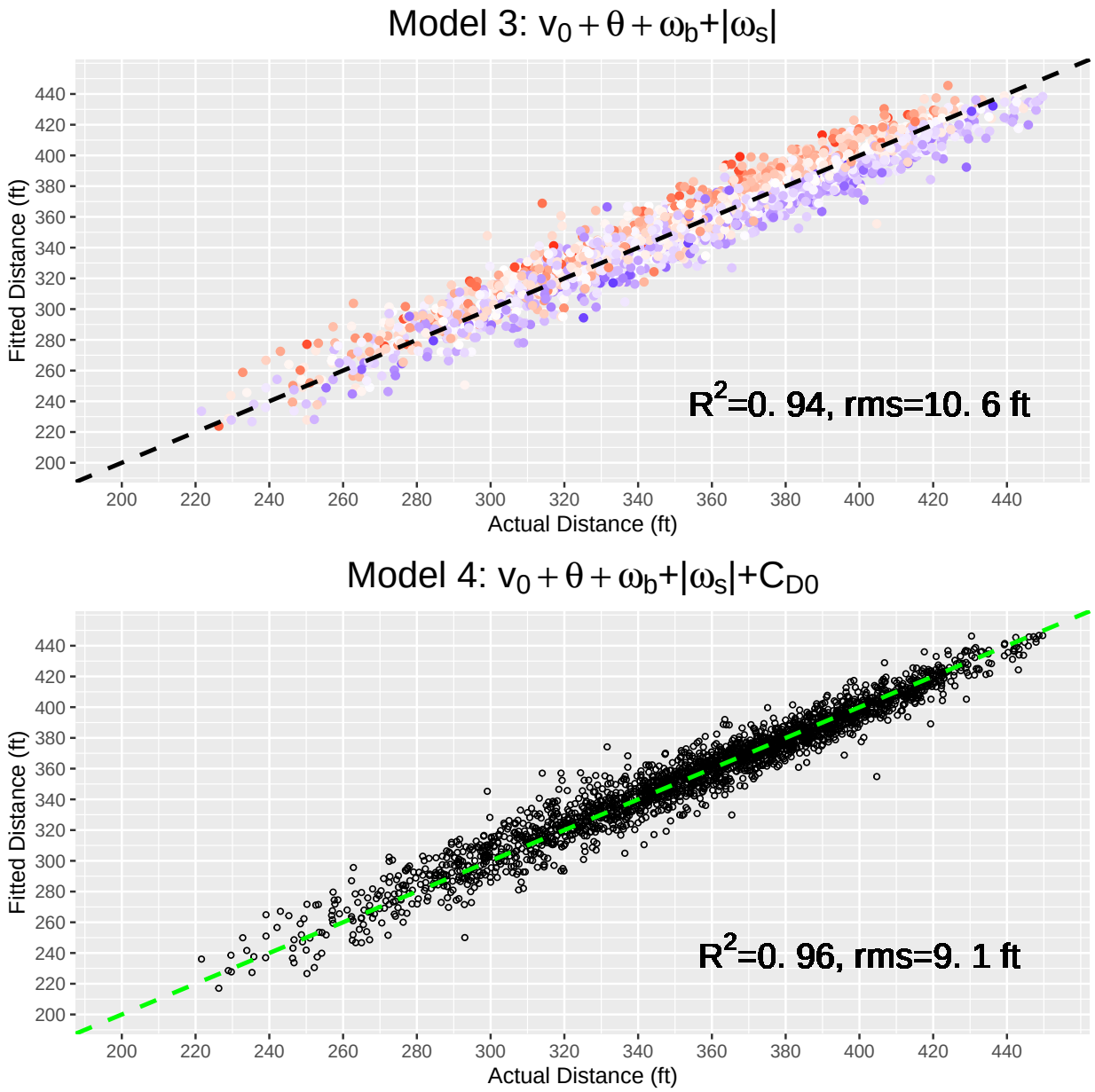

FIG. 1: Plot of fitted vs. actual distance for Models 3 and 4, as indicated in each graph, with the dashed line representing equality. The square of the Pearson correlation coefficient R^2 and the rms deviation of the data from the fit are indicated. In the top graph, in which all physically relevant variables other than C_{D0} are included, the colors indicate C_{D0}, with blue the largest (~ 0.34), white the midrange (~ 0.30), and red the smallest (~ 0.26), and clearly show the inverse relationship of distance to drag.

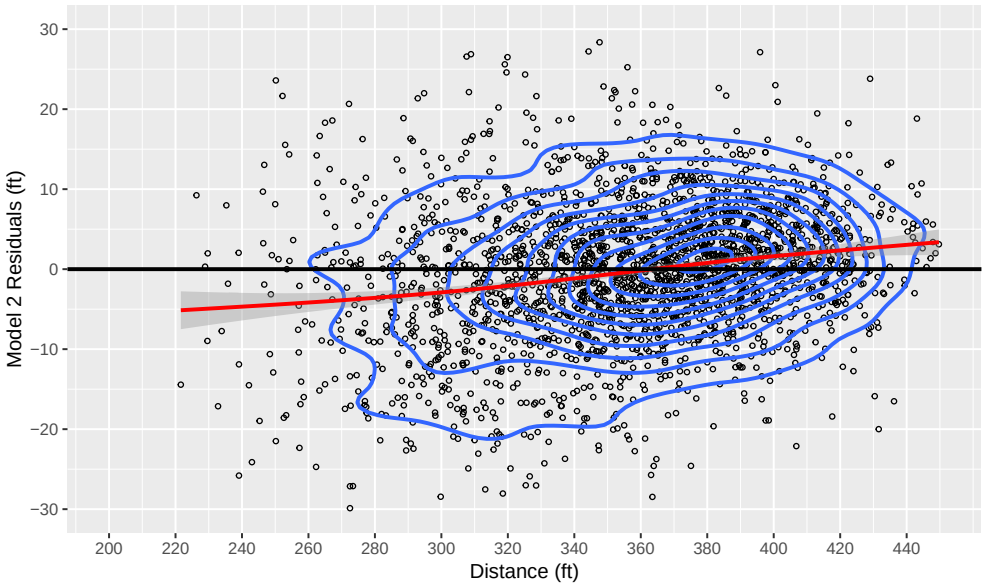

FIG. 2: Plot of Model 4 residuals vs. actual distance, along with a trend line. The tilting of the latter shows a bias in the fit. Nevertheless, the model is accurate to ±2 ft over the range 320-420 ft.

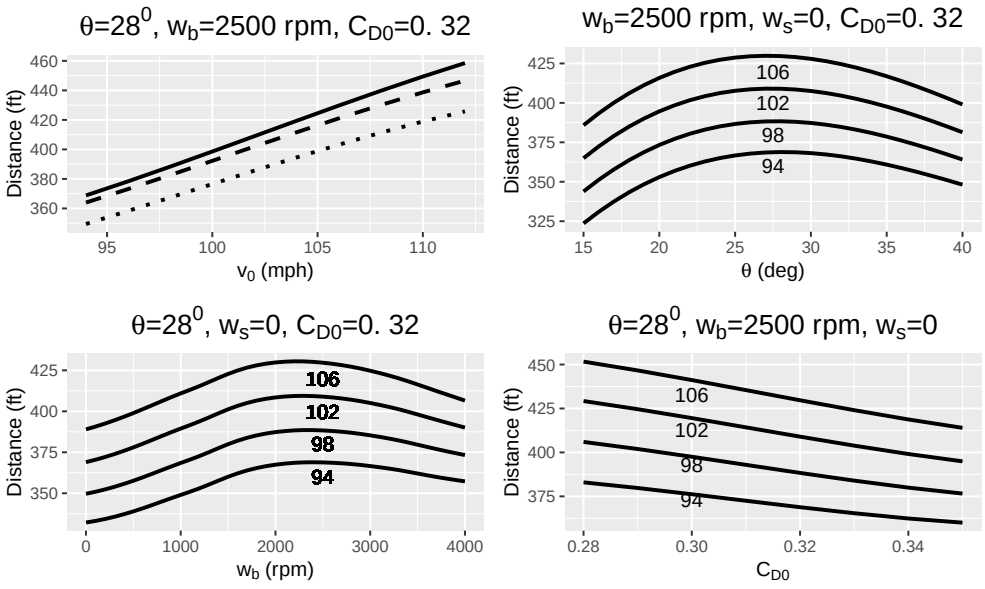

FIG. 3: Model 4 fits of batted-ball distances, with fixed parameters indicated on the graphs. Upper left: Distance vs. exit velocity for $\omega_s = 0$ (solid), ω_s=1000 rpm (dashed), and ω_s=2000 rpm (dotted). Upper right: Distance versus launch angle for various exit velocities. Lower left: Distance vs. backspin for various exit velocities. Lower right: Distance vs. C_{D0} for various exit velocities.

IV. CONTRIBUTIONS TO THE MEASUREMENT NOISE

From Table I, one can determine the contribution of the various parameters to the variance in batted-ball distance for fixed exit velocity and launch angle. These results are shown in Table II. By far, the largest contribution are the individual spin components, which are determined from the combination of total spin rate and spin axis. The contribution of C_{D0} is similar to that found in the earlier study and is interpreted to be the variation in distance due to a ball-to-ball variation in the drag coefficient. But what about the random measurement noise, which is actually somewhat larger here than earlier, quite possibly due to the larger launch angle range? To what do we attribute that? That is the focus of this section.

Under the assumption that the measured launch parameters (v_0, θ, $\omega_{\rm b}$, and $\omega_{\rm s}$) are relatively noise-free,[3] the two remaining sources of noise are in the measurements of the distance d and the spin-independent drag coefficient C_{D0}. Given that the total contribution of noise is 9.1 ft (Table II), the question I ask is how is that noise distributed between d and C_{D0}? I start by defining the quantities of interest:

- σ_b = contribution to variation in distance due to ball-to-ball variation in C_{D0}. This quantity was determined in Model 4 and is listed in Table II. In this section, I take an alternate approach.

- σ_m = contribution to the random measurement noise due to the measurement of C_D.

- σ_d = contribution to the random measurement noise due to the measurement of d.

I start with Model 3, which includes all physical parameters other than C_{D0}, and investigate how the residuals depend on C_{D0}. This is shown in the scatter plot of Fig. 4. The dashed line in that plot is a simple linear regression. However, that procedure assumes that any noise is entirely in the dependent variable (d) and there is none in the independent quantity (i.e., σ_m=0)). When that assumption doesn't hold, i.e., when there is random variation in both quantities, then a different procedure must be used, the so-called Deming regression.

[3] While I am confident in this assumption, it is clearly speculation on my part. In any case, let's see where it leads.

To utilize this procedure requires knowledge of the ratio $r = \sigma_m/\sigma_d$, which is not *a priori* known. Indeed, it is exactly what we are trying to find. My approach is an iterative one. First, I note that the rms of the Model 3 residuals is the Pythagorean sum of σ_b and σ_d.[4] Next I assume a ratio r and perform the Deming regression to find the slope S of d-vs.-C_{D0}. The rms of the residuals is the Pythagorean sum of σ_m and σ_d. Finally, the standard deviation of the C_{D0} distribution, after multiplying by the slope S, is the Pythagorean sum of σ_b and σ_m. These three relationships can be solved to find the three quantities of interest, from which r can be found and compared to the starting value. This process is iterated until the final value agrees with the initial value. The results are shown in Table III.

The ball-to-ball contribution, σ_b, is slightly larger but generally agrees with the value found from Model 5 (see Table II). Further, we see that the random measurement noise is completely dominated by the distance measurement, with only nominal contribution from the C_D measurement. Another way to see the same thing is shown by the curves in Fig. 4, where the simple linear regression line is not all that different from the Deming regression line, confirming that the contribution of the C_D measurement to the noise is small compared to that of the d measurement. A comparison between the Model 4 and 5 fits is given in Fig. 5. They are clearly similar, although the Model 4 fit is marginally better. I am satisfied that the question I posed has been answered.

TABLE II: Contributions to the variance in batted-ball distance for fixed exit velocity and launch angle.

Parameter	rms (ft)	fraction
ω	6.1	6.5%
$\omega_{\rm b}$, $\omega_{\rm s}$	20.5	73.8%
C_{D0}	5.4	5.1%
noise	9.1	14.5%
total	23.9	100%

[4] By Pythagorean sum, I mean $\sqrt{\sigma_b^2 + \sigma_d^2}$.

TABLE III: Results from Model 5.

Contribution	Parameter	rms (ft)
ball-to-ball variation in C_D	σ_b	5.7
random measurement noise, C_D	σ_m	3.1
random measurement noise, d	σ_d	8.9

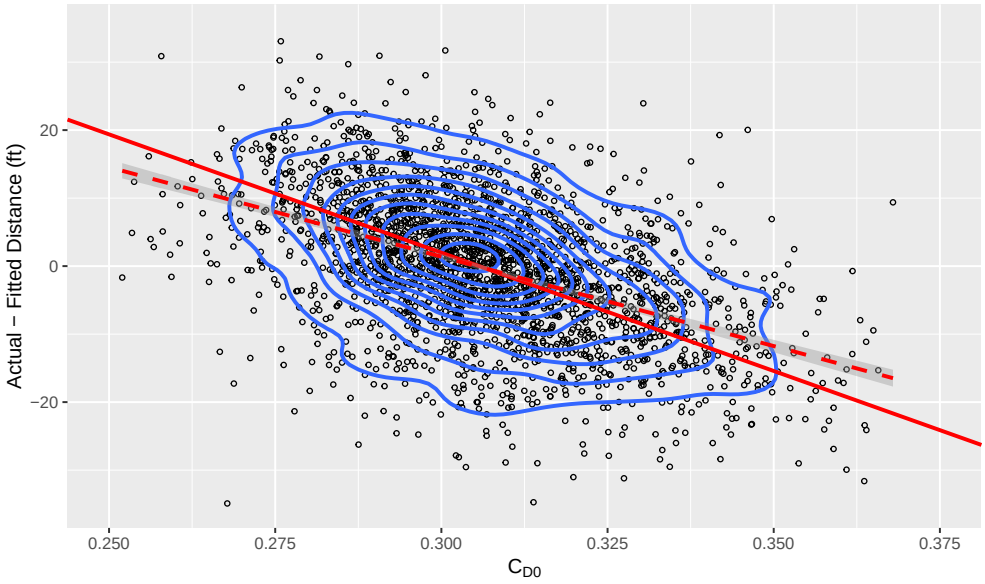

FIG. 4: Scatter and density contour plot of the Model 3 residuals vs. C_{D0}. The dashed line is the result of a simple linear regression while the solid line is the result of a Deming linear regression, as discussed in the text.

V. COMPARING C_{D0} FROM PITCHED AND BATTED BALL

I now come to the final item on my agenda, a comparison of C_D values for the batted ball and the corresponding pitched ball. Care is necessary in this process, since we know C_D depends on spin; in fact, it depends on the "transverse spin", the component of spin perpendicular to the velocity vector. For the batted ball, models of the ball-bat collision show that the spin of the ball just after leaving the bat is almost completely transverse and composed of a combination of backspin and sidespin. But that is definitely not true of the

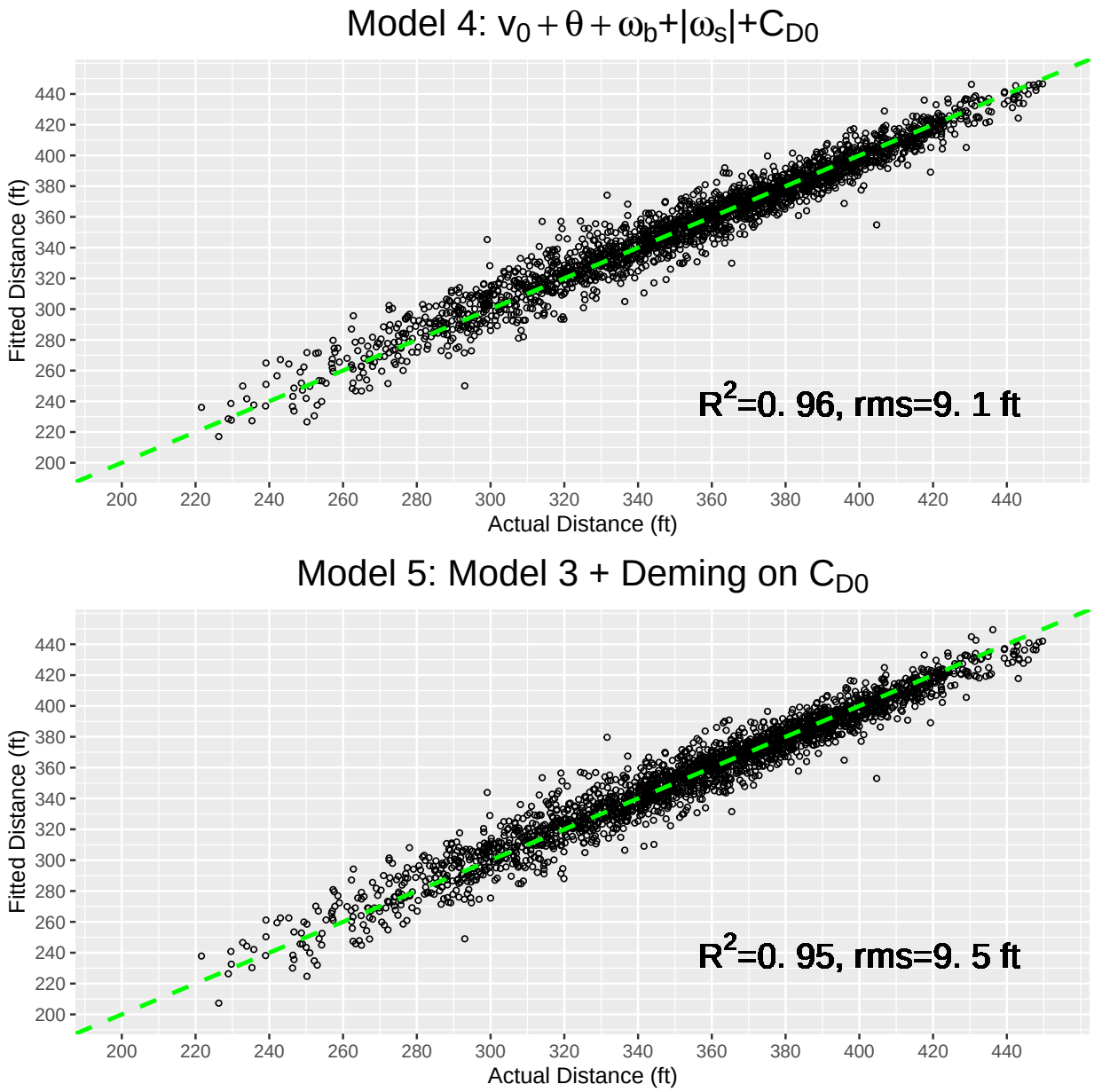

FIG. 5: Plot of fitted vs. actual distance for the two different models that control for C_{D0}, as indicated in each graph, with the dashed line representing equality. The top plot is from Model 4 and is identical to the bottom plot in Fig. 1. The bottom plot is the result of the Deming linear regression of the residuals of Model 3 to C_{D0}. The rms deviation of the data from the fit are indicated.

pitched ball, as my former student Charlie Young and I discussed in a recent article.[5] The procedure for obtaining both the drag coefficient and the transverse spin is discussed in yet

[5] In the article, the notation "active spin" is used instead of "transverse spin".

another unpublished article, from which the spin-independent C_D value can be obtained. A scatter plot of the pitched-ball and batted-ball C_{D0} values is presented in Fig. 6. The values are genearlly in agreement, albeit with considerable scatter (R^2=0.147).

One might argue that it is better to use the pitched-ball C_{D0} values when analyzing the distance data since the former is completely independent of the latter. And I agree with that argument as a matter of principle. However, in practice that is not so feasible, since the pitched-ball values have considerable measurement noise, in part due to the direct measurement of C_D and in part due to the determination of the transverse spin, both of which are measured over a relative short flight path. Indeed, when I use the pitched-ball values in Model 4, the rms of the residuals are barely an improvement over Model 3, which did not include C_{D0} as a parameter.

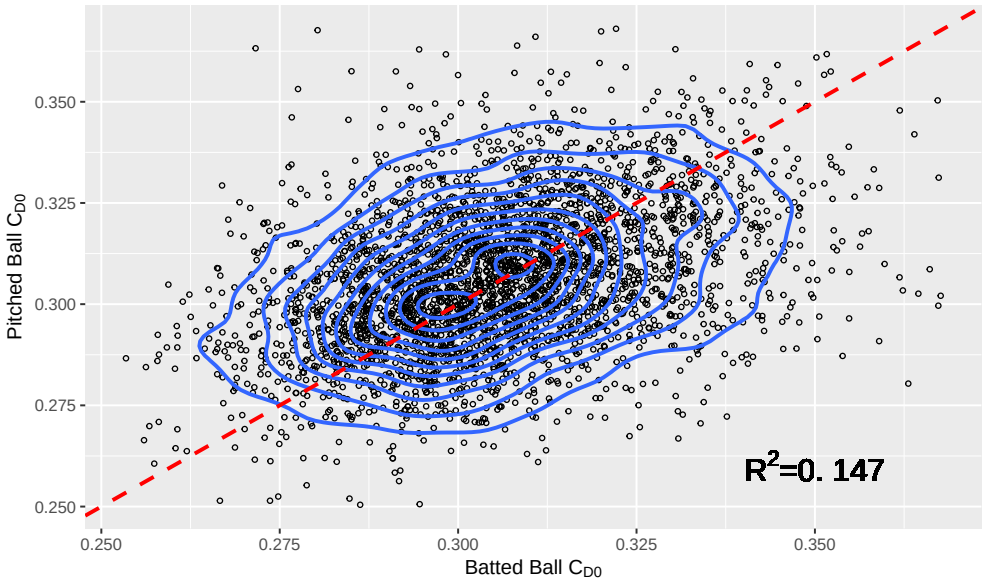

FIG. 6: Scatter and density contour plot of C_{D0} values for the pitched and batted balls. The solid line has zero intercept and unit slope and represents equality. The square of the Pearson correlation coefficient, R^2, is given.

VI. SUMMARY

In this article, I have re-visted the question of variation of fly ball distance, using what I regard as an improved model applied to a much large data set. By randoming splitting the data into two halves, I have separated the issue of training the model from that of

testing the model. I have drilled down on the contribution of measurement noise, showing that it is dominated by random noise on the distance measurement, with only a small contribution from the random noise on the C_D measurement. Finally, I have shown that the C_{D0} measurements on the batted ball and corresponding pitched ball are correlated, as they should be, albeit with a lot of scatter in the data.