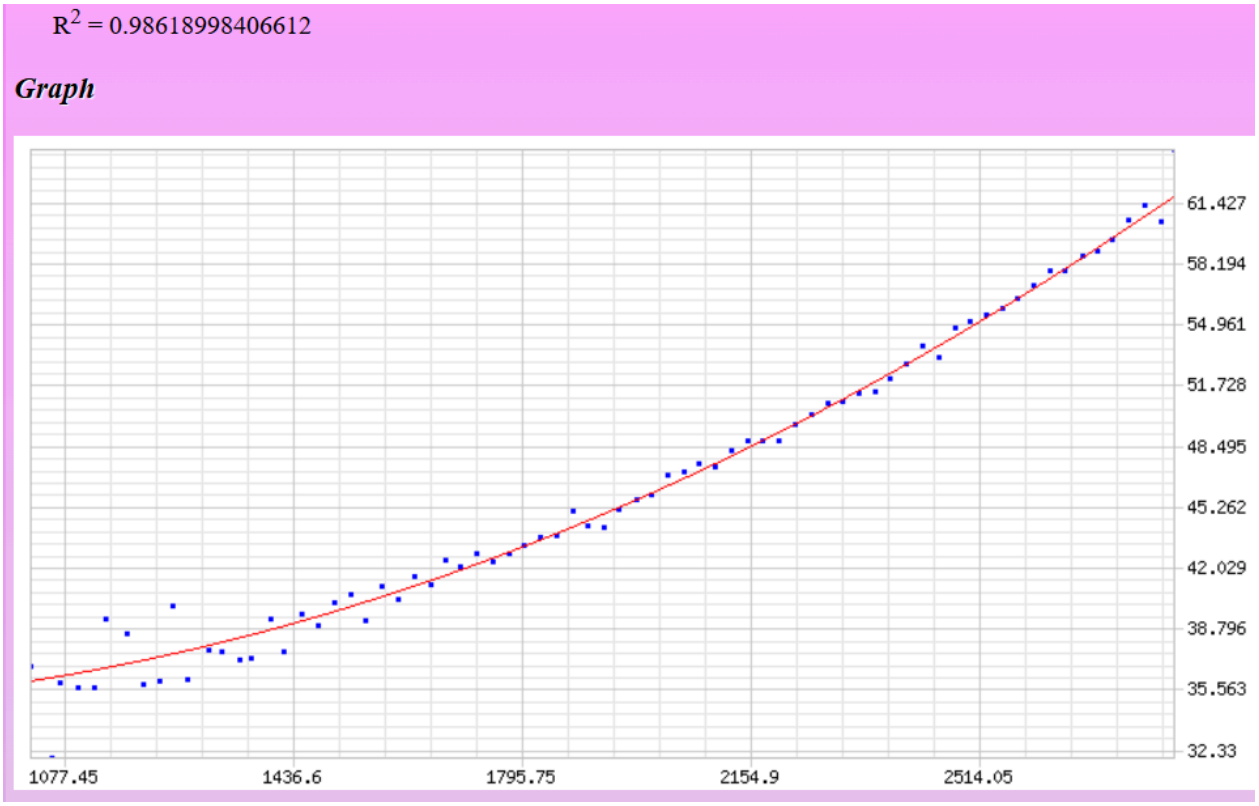
CSE702 Week 2+: Elo Ratings and Raw Metrics

We will focus on **T1** match (called MMP for move-match % in my code), **ACPL** (average centipawn loss without scaling), and **ASD** (scaling ACPL according to the position value).

Toward the end of the 2010s, as more games by FIDE-rated players under 1600 became available, I noticed that the graph of T1-match versus Elo rating was no longer best explained as a straight line. For continuity of settings, rather than graphs using older versions of Stockfish and other engines in Multi-PV mode, here are graphs using Stockfish 16 in Single-PV mode for the in-person Chess Olympiads in 2010, 2012, 2014, 2016, and 2018. All regressions and diagrams are from Dr. Andrew Que's Polynomial Regression applet, and all of the data to paste into the applet is annotated in this downloadable spreadsheet.

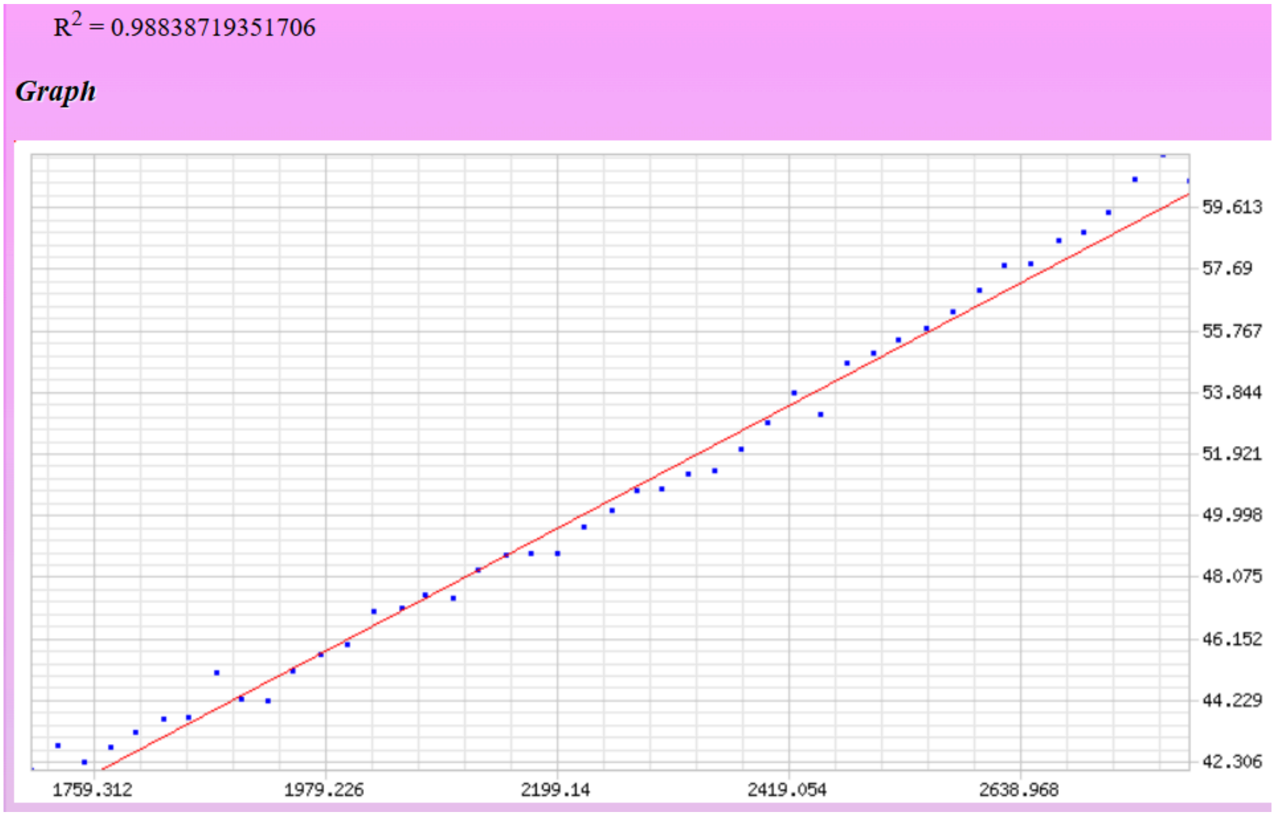
The graph of Elo to T1-match in 2010--2018 is definitely not a straight line:



$R^2 = 0.98618998406612$

Graph

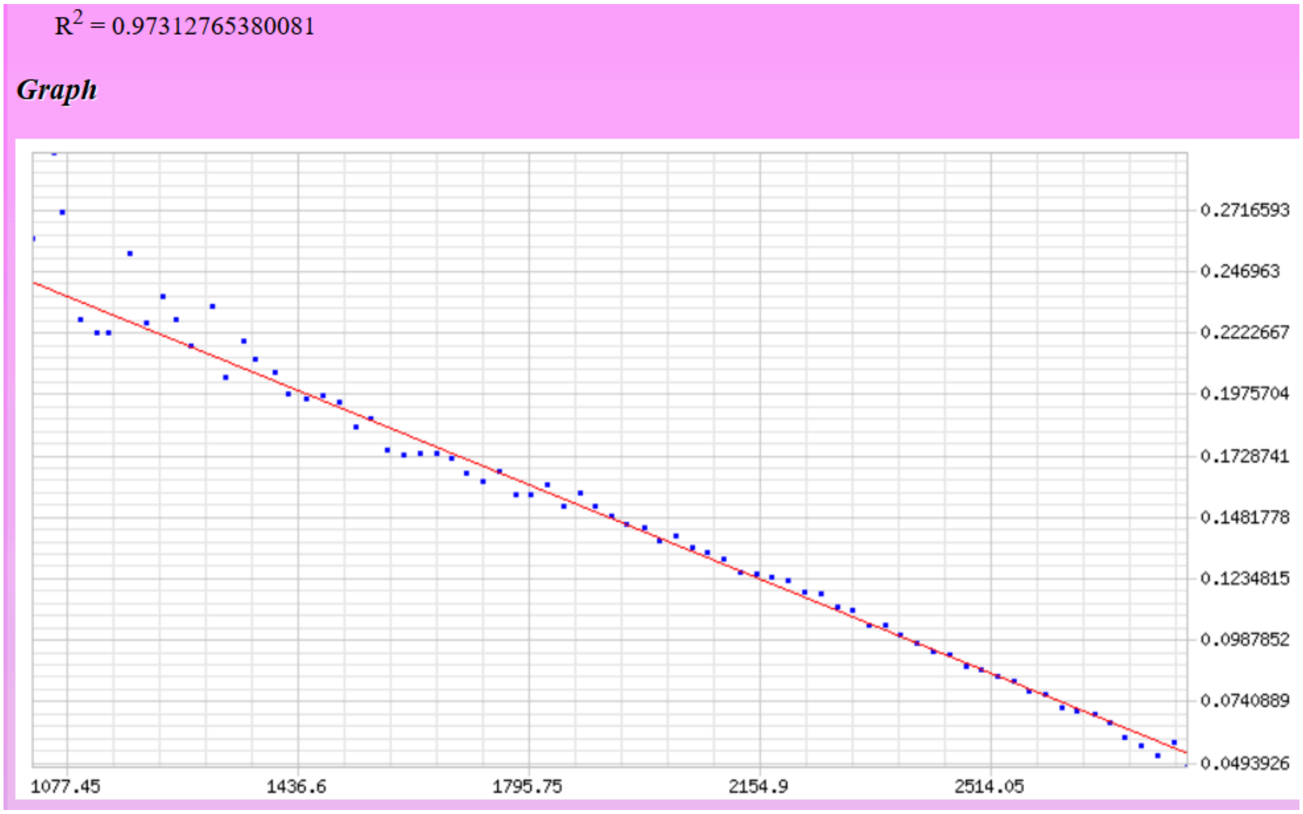
Scrubbing data below 1700 and the point for bucket-averaging-2820 at upper right makes a straight line highly believable, however:



$R^2 = 0.98838719351706$

Graph

Moreover, the unscrubbed graph for ASD is a straight line clear down under 1400:

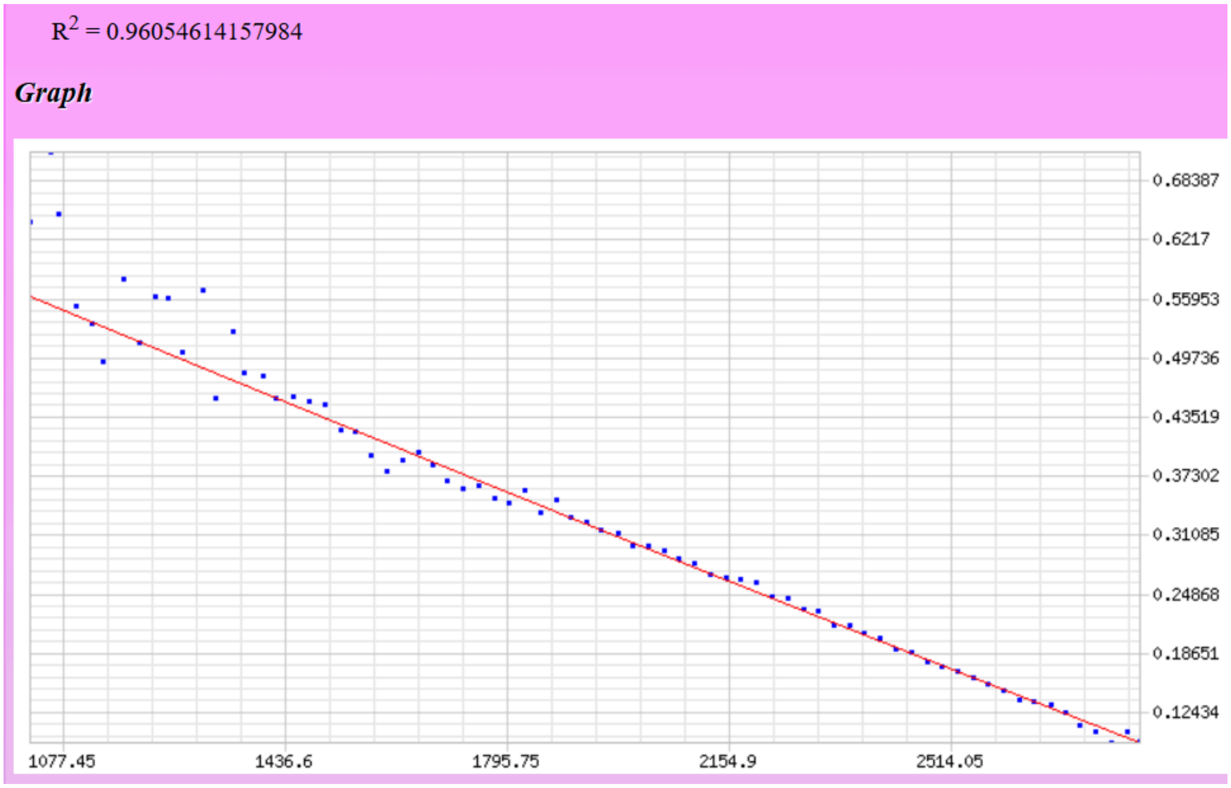


$R^2 = 0.97312765380081$

Graph

This could be an artifact of the formula used for the scaling. In Single-PV mode I use a "one-size-fits-all" logarithmic formula whose imperfection shows even for 1800 vis-a-vis 2200 in my 2016 "When Data Serves Turkey" article. The same graph for unscaled ACPL is highly similar, however.

ACPL (no scaling but blunder magnitude capped at 4.00):

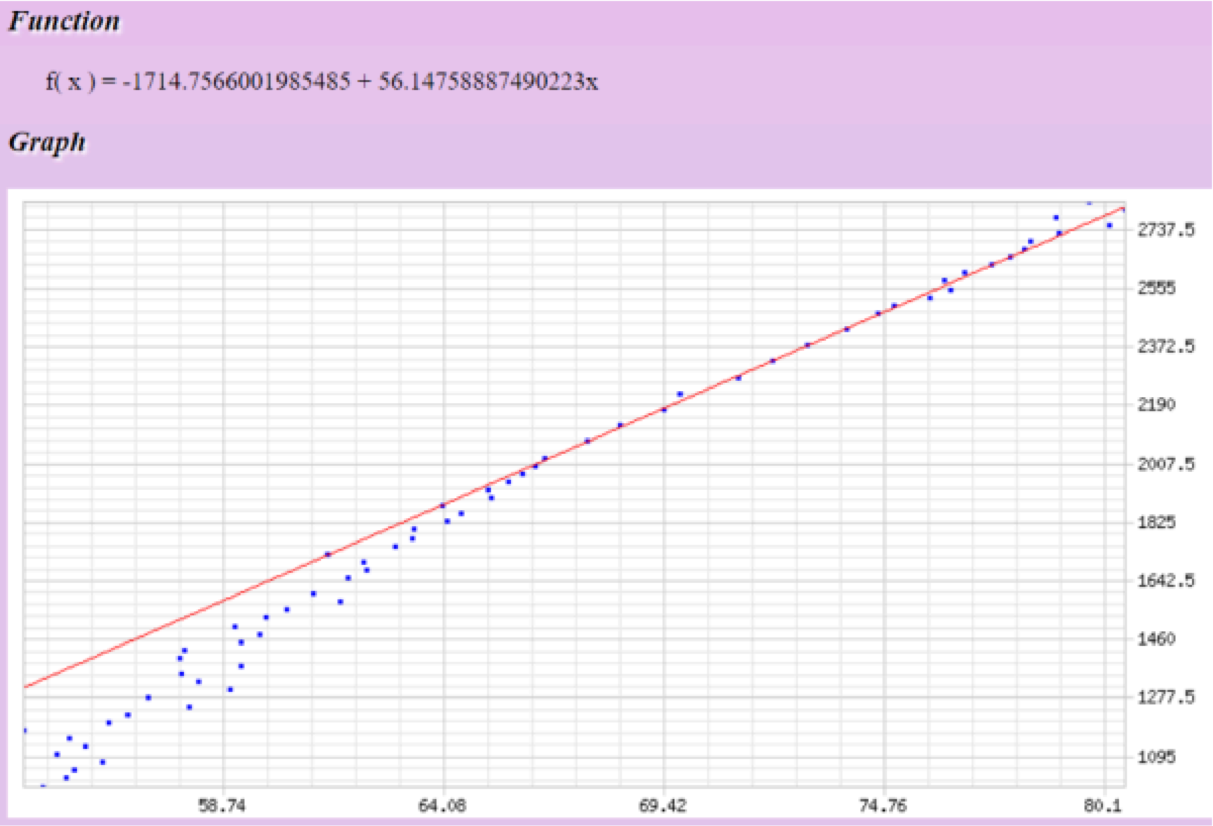


$R^2 = 0.96054614157984$

Graph

My "Raw Outlier Index" (ROI) for Single-PV mode, which combined T1-match and ASD, wound up being a cubic regression---which as noted in the 2019 article gives more-plausible extrapolation. And for the IPR curve of my full model, I wound up using a nonic (9th-power) regression for better control of the extrapolation down to Elo rating 100 and tighter accuracy elsewhere. Elsewhere in my full model, I used cubic for the main model parameters, though those too are close to straight lines---as shown in the chart at upper left here.

So in late 2019 I chalked the issues up to ignorance about ratings below 1500, for which I calibrated *ad-hoc* and double-checked my model in copious scholastic online tournaments that were held during the pandemic. In 2023, FIDE statistician Jeff Sonas published analysis that lifted my ignorance. My support for his proposal to map the interval [1000,2000] linearly onto [1400,2000] was clinched by an update to my August 2023 article showing the following graph of the "**T3-0.5**" metric, which counts a match if the played move is in the engine's top 3 and no more than a 50-centipawn error overall, against Elo in 2010--2019:



Function

f(x) = -1714.7566001985485 + 56.14758887490223x

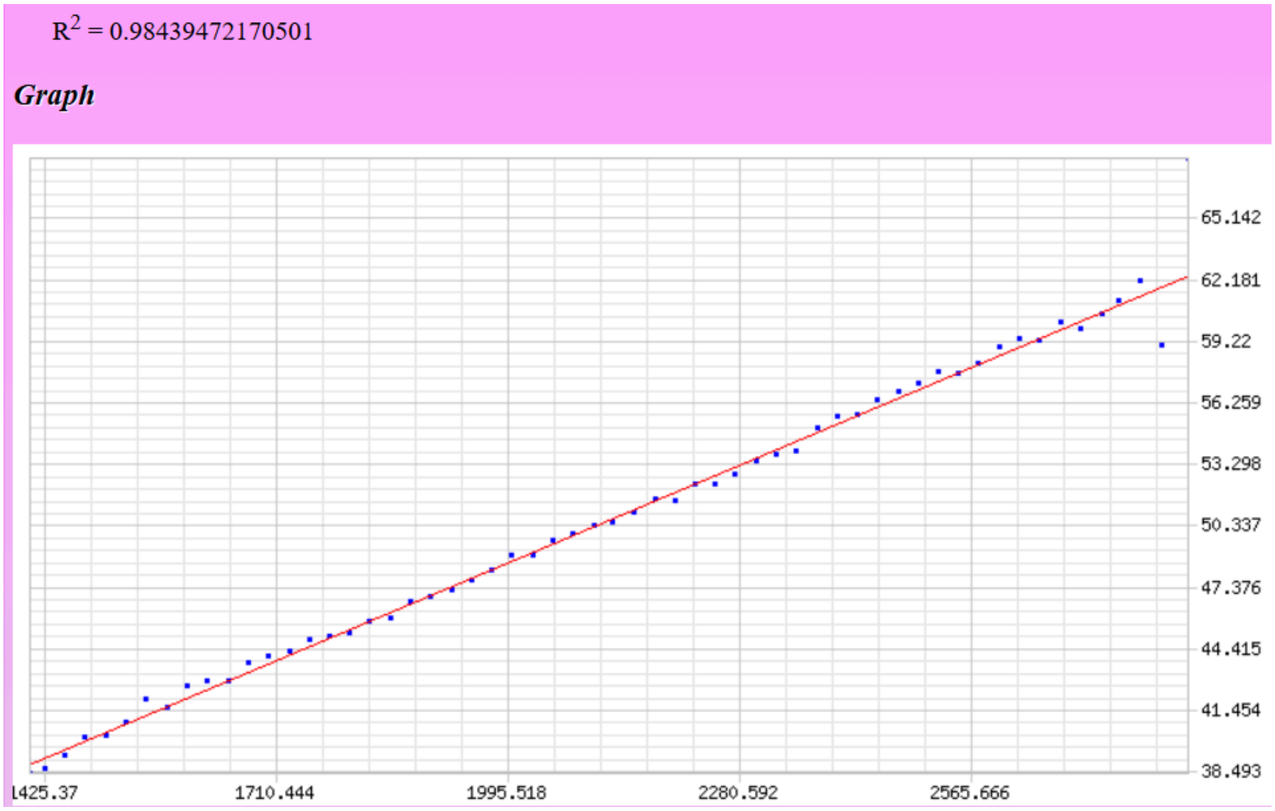
Graph

This looks more like a line that is kinked right near 2000 than a quadratic or cubic curve, and it is almost exactly fixed to linear by the Sonas mapping. FIDE made the change effective on the March 2024 rating list (released on March 1). How has it been going?

T3-0.50 can only be done in Multi-PV mode, so we will stay with T1 and ASD for the most part. The 2020 in-person Olympiad was canceled, and 2022 was held in Chennai at the apex of "pandemic lag" in official ratings. With or without my rating adjustments, T1 still comes out curved and ASD is believably a line.

July--December 2024 Results

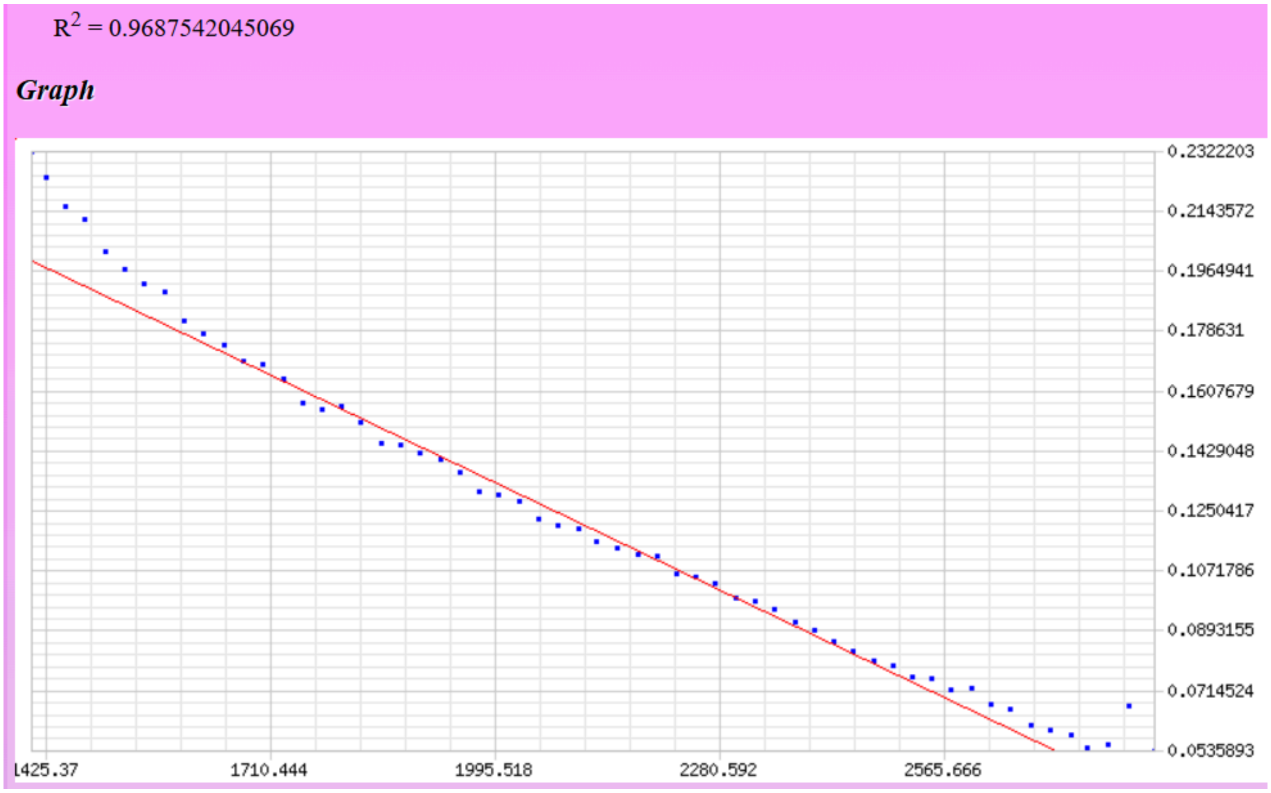
T1 match, no scrubbing---note that the floor is now 1400:



$R^2 = 0.98439472170501$

Graph

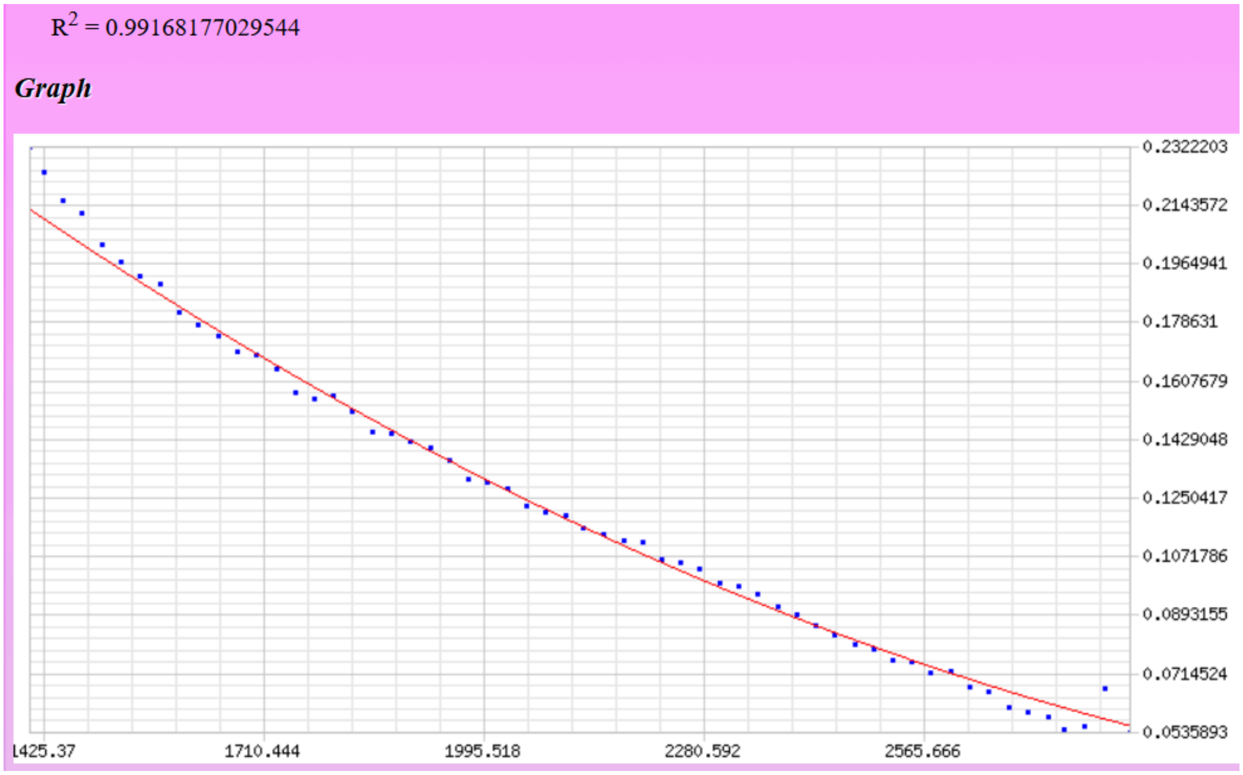
Impressively linear. But for ASD, no scrubbing:



$R^2 = 0.9687542045069$

Graph

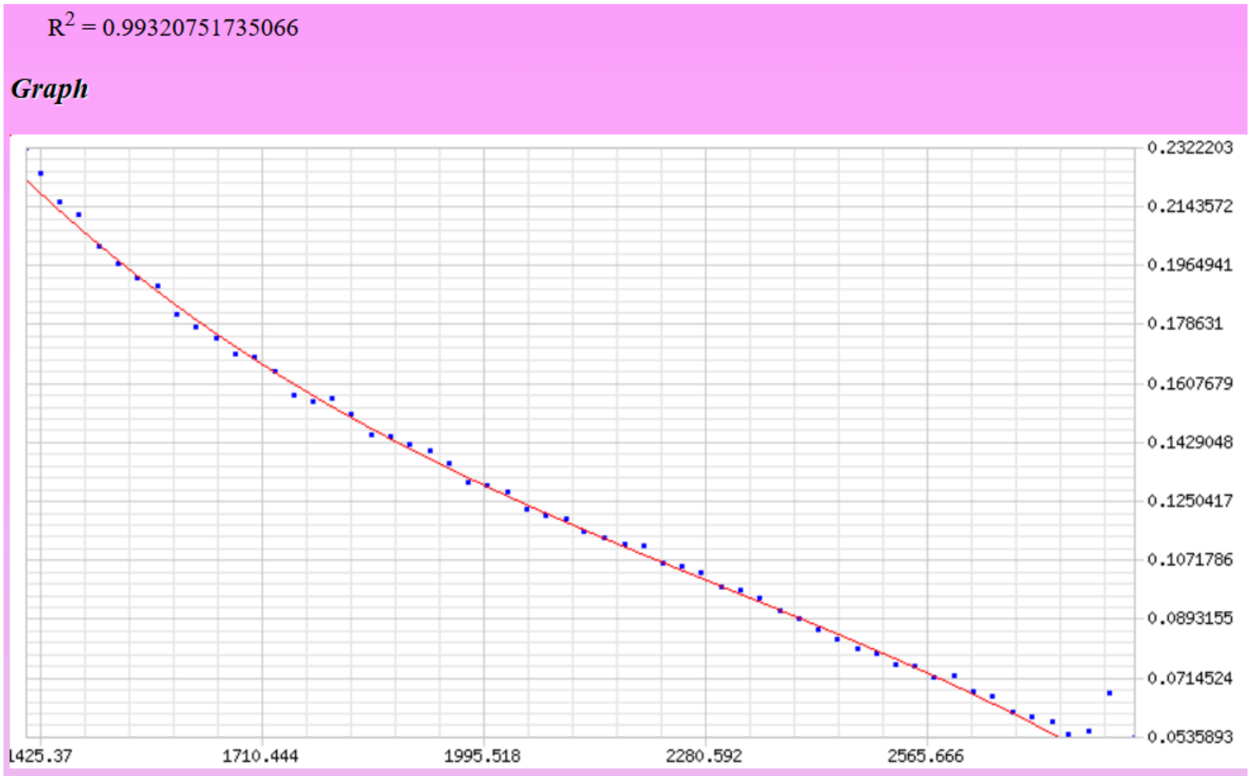
This is definitely not a line. The quadratic fit is much better:



R^2 = 0.99168177029544

Graph

A cubic fit is even better at the tail:



R^2 = 0.99320751735066

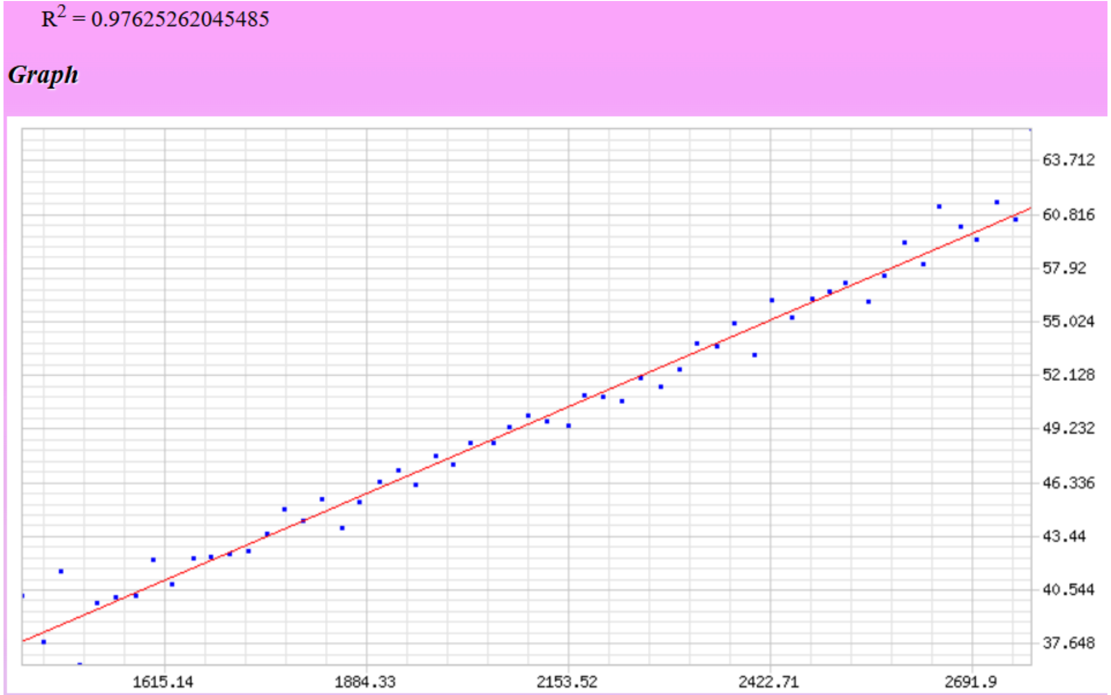
Graph

When the three readings at right for highest Elo are scrubbed the cubic fit gives R^2 almost 0.998. But another possibility is that ratings of the lowest players have been raised too far. If we map [1400,2000] back to [1200,2000] the graph straightens out considerably.

Data from September 2024 Olympiad

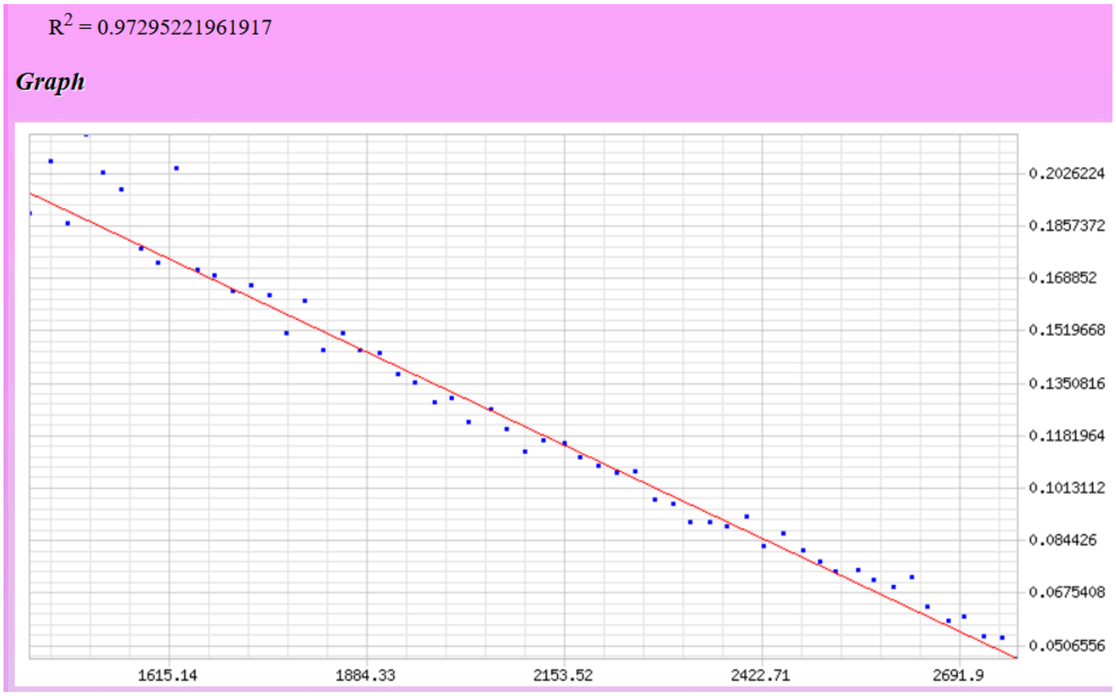
The Budapest Olympiad in September 2024 gave six months for ratings to adjust. How then?

T1-match:



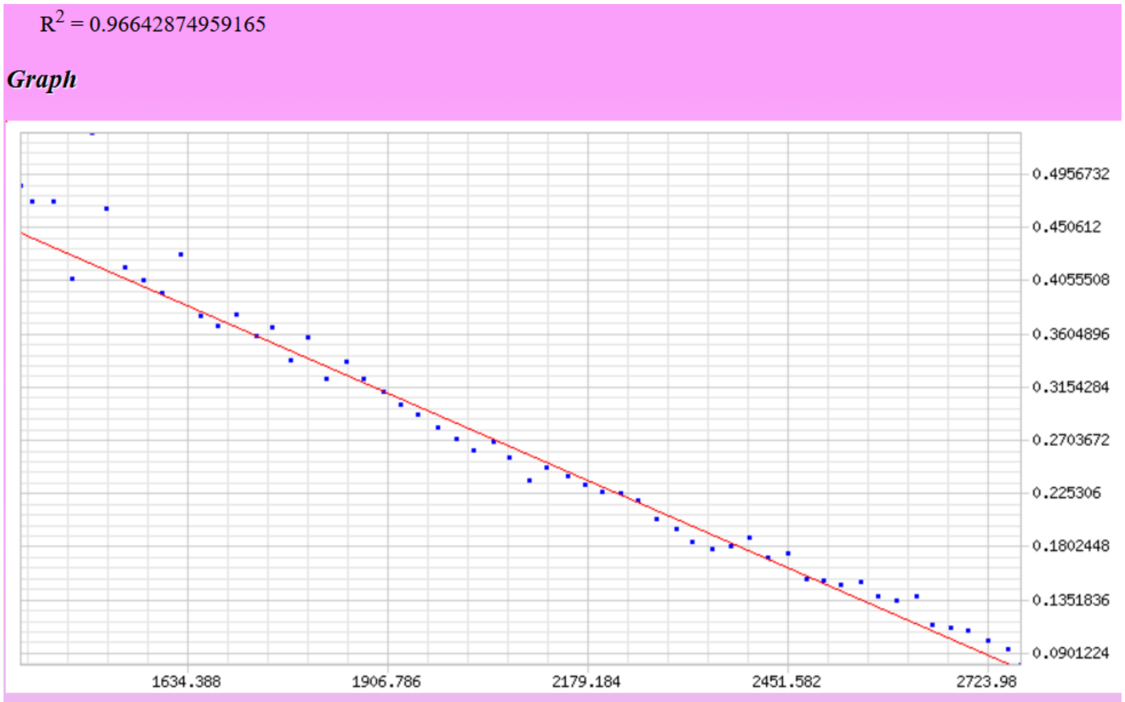
$R^2 = 0.97625262045485$

A beautiful line. Even better when the 65%-matching point for the top bucket averaging 2770 is scrubbed---the R^2 correlation measure becomes 0.981. For ASD:



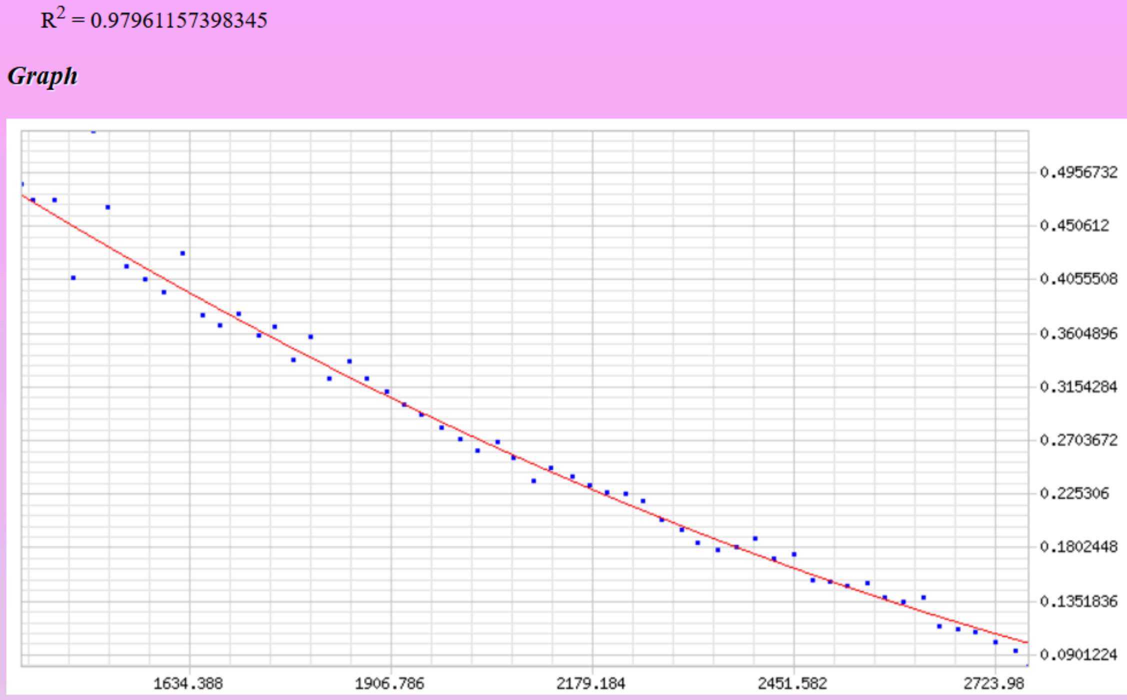
$R^2 = 0.97295221961917$

Also good---though with noise below 1625 rating that is mostly above the line. Put another way, the formula $asd = 0.3549 - 0.0001114 * Elo$ gives $asd = 0.1989$ at rating 1400, basically 0.20. This means that any ASD figure above 0.20 corresponds to ratings below the current FIDE floor. For ACPL, the corresponding number is about 0.45 from the line---


$R^2 = 0.96642874959165$

Graph

---but is is over 0.50 from the data, which fits a quadratic curve noticeably better:


$R^2 = 0.97961157398345$

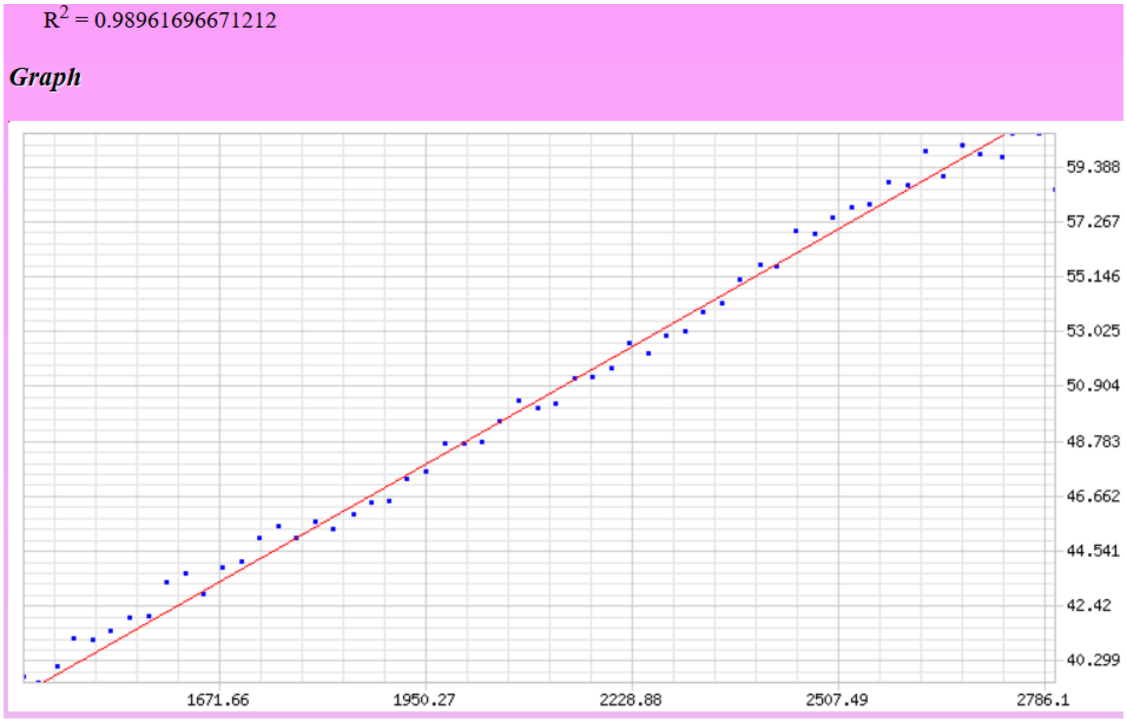
Graph

But OK, we have both T1 and ASD behaving well at the Olympiad. How about since?

Data From Oct.--Dec. 2024:

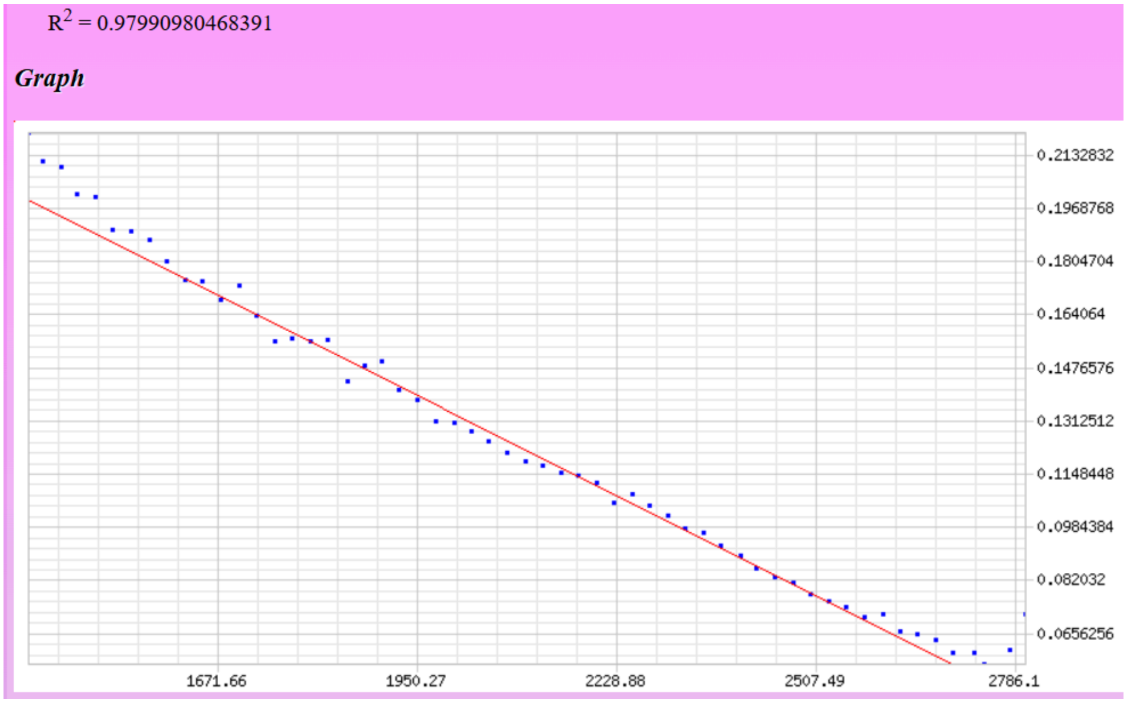
There are three tiny buckets under 1400: 234 moves averaging 1152 Elo, 71 moves averaging 1200, and 280 moves averaging 1352, compared to 2,156 moves for the bucket averaging 1407 rating and 1,981 moves for 2800 at the top. Scrubbing the three buckets gives:

T1-match:


$R^2 = 0.98961696671212$

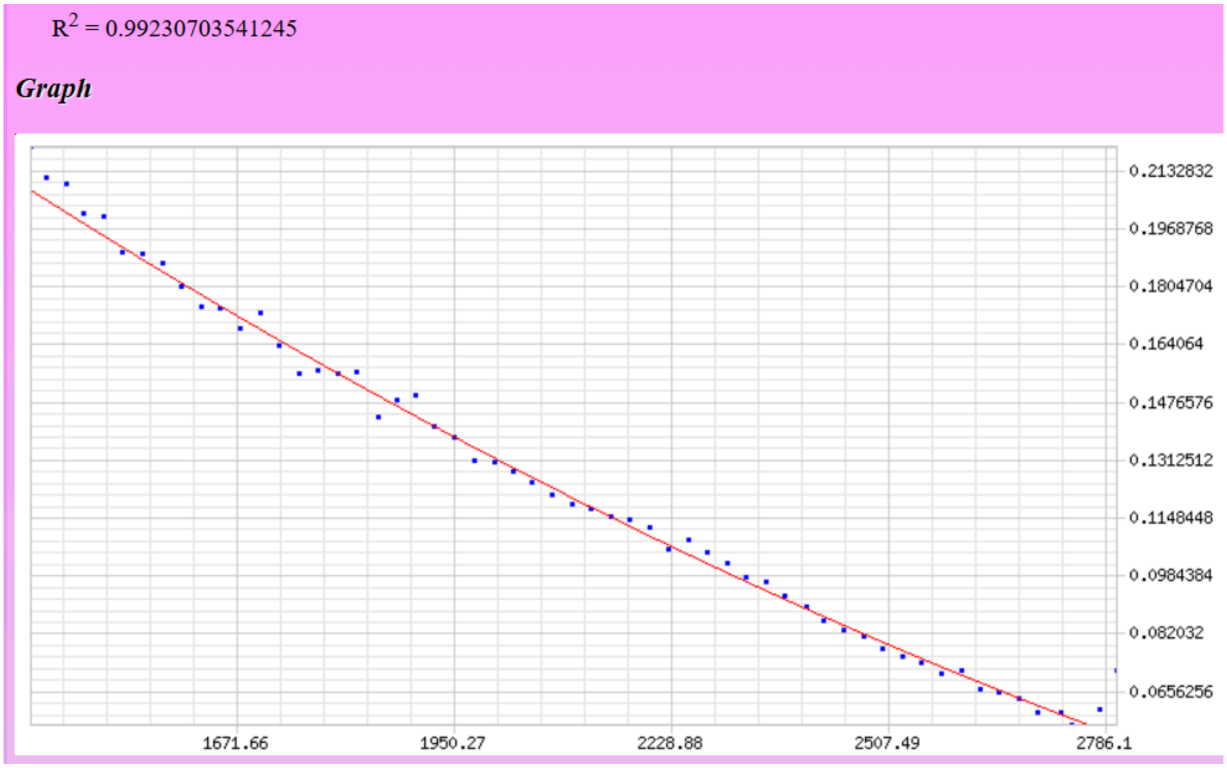
Graph

A very good straight line. But for ASD---


$R^2 = 0.97990980468391$

Graph

---it is better as a quadratic fit:



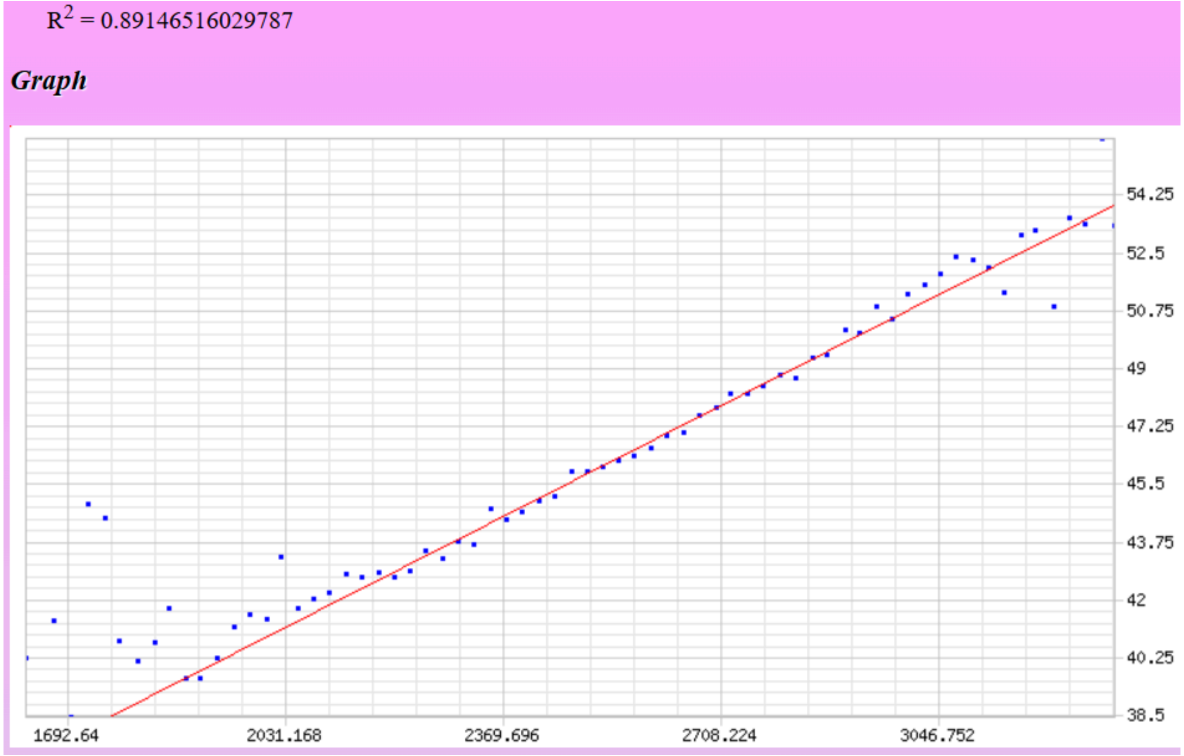
$R^2 = 0.99230703541245$

Graph

Chess.com Titled Tuesday, Oct.--Dec.

It is interesting to compare with Chess.com's rating system as used for their weekly "Titled Tuesday" online blitz tournaments. This is on a different scale from FIDE, though it uses the same logistic formula to figure points expectation from a difference in ratings. The top ratings are in the 3300s. Again using Stockfish 16 for October through December:

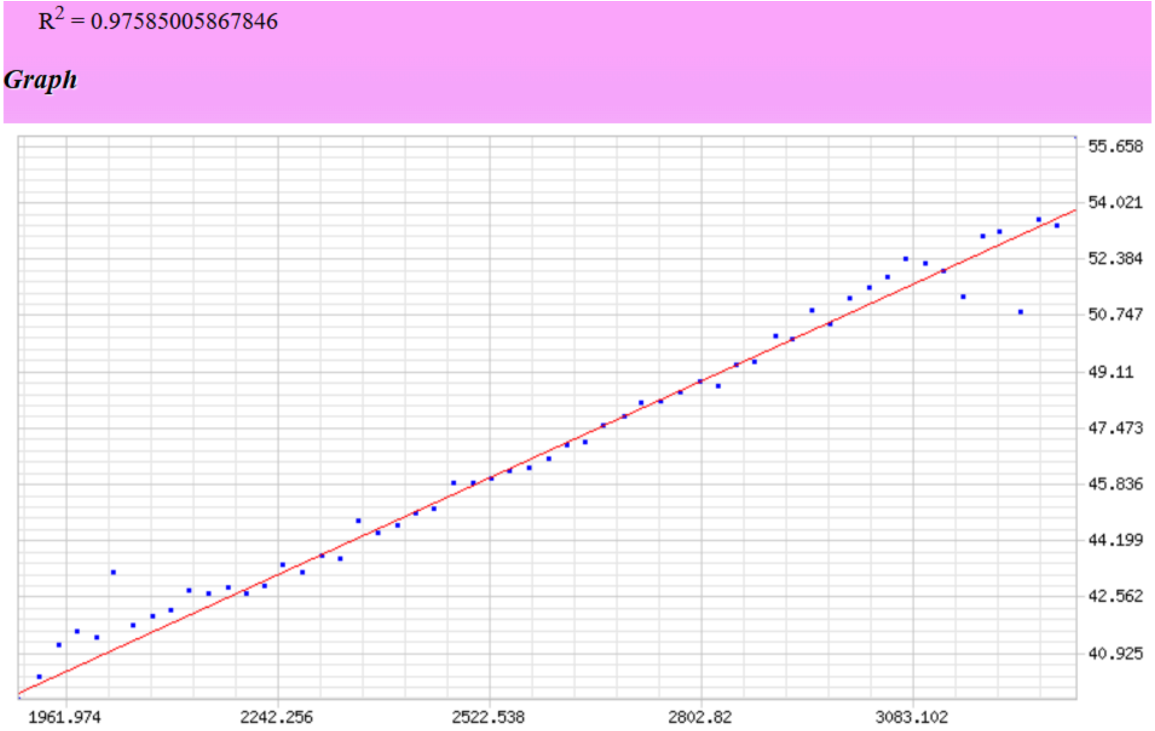
T1-match:



$R^2 = 0.89146516029787$

Graph

This could be better as a quadratic. But the buckets under 1900 rating and the top bucket are small. Scrubbing them leaves

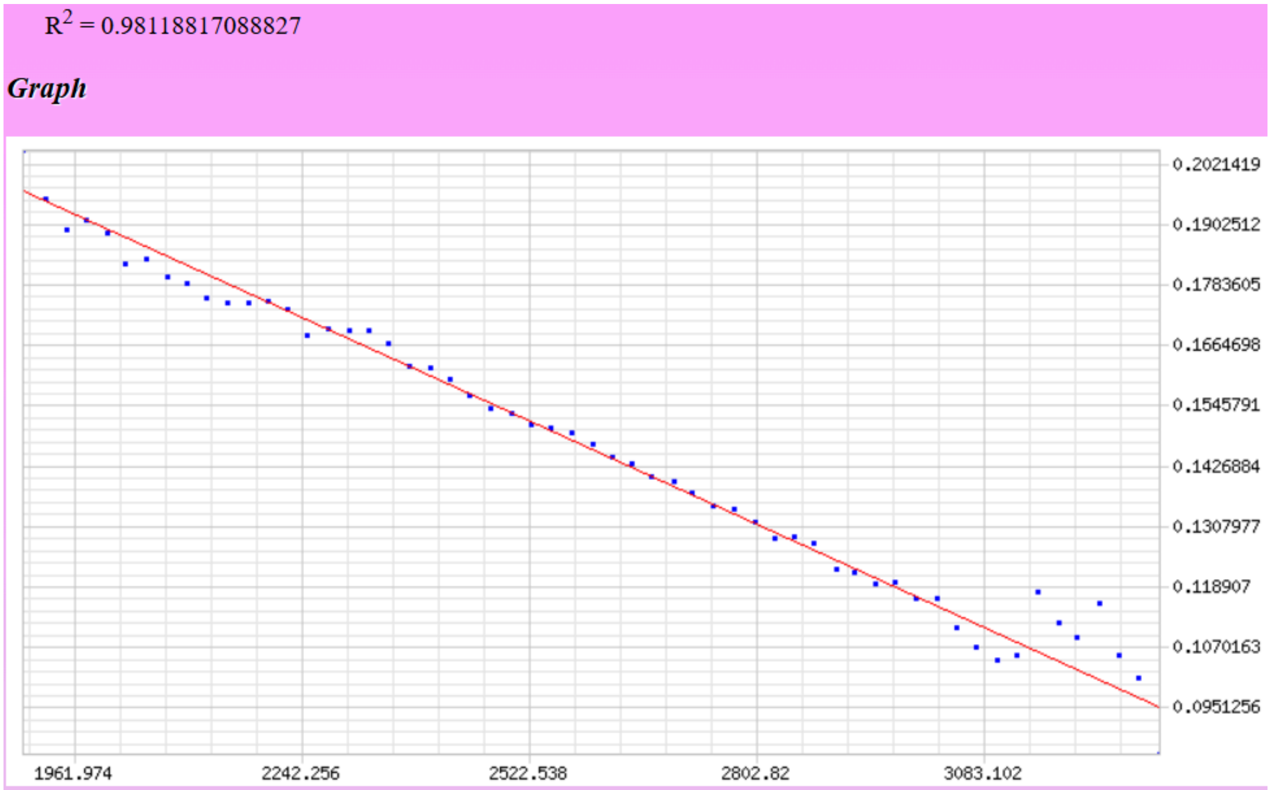


$R^2 = 0.97585005867846$

Graph

which is a perfectly fine line.

ASD, same scrubbing:

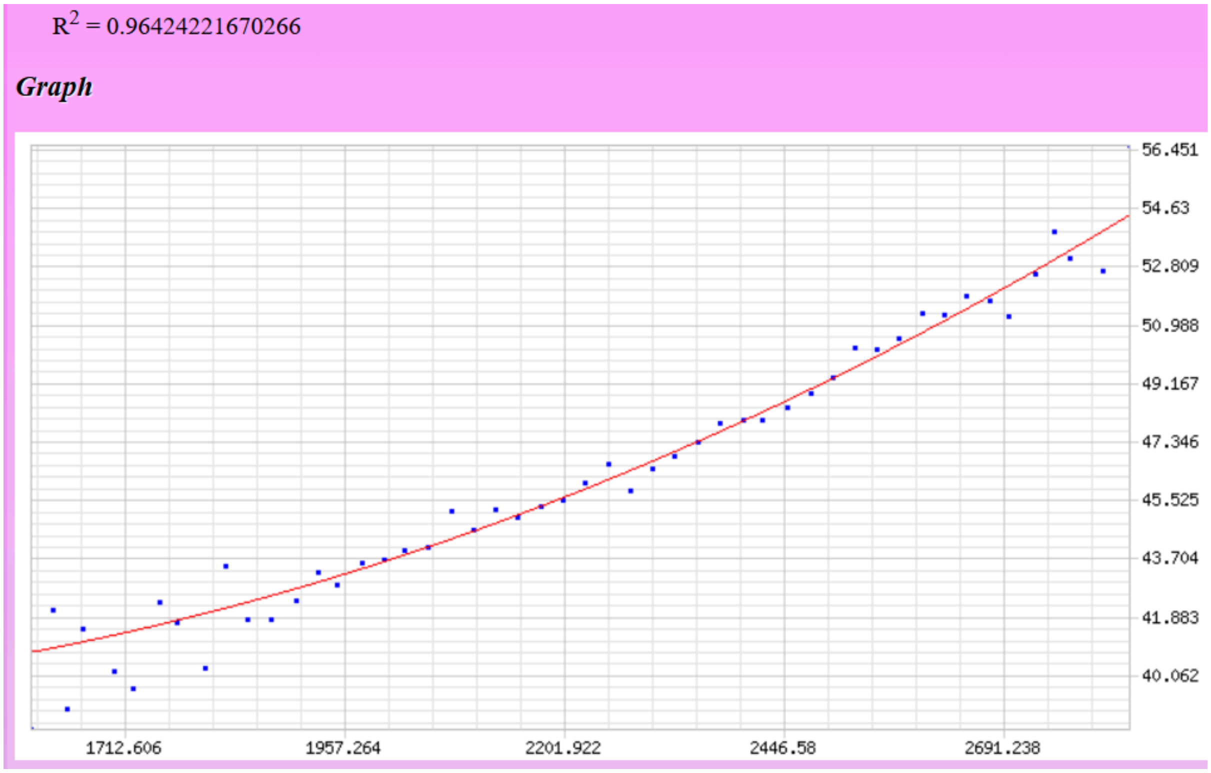

$R^2 = 0.98118817088827$

Graph

Again, a perfectly good line.

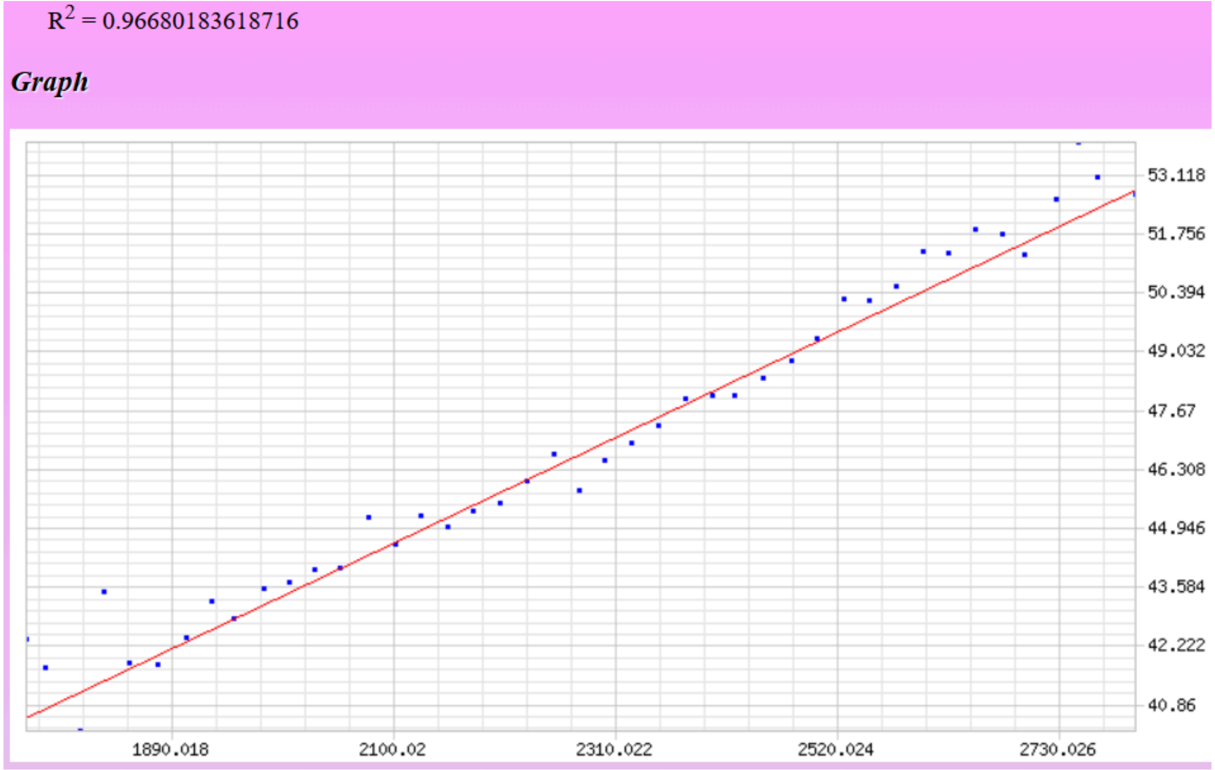
It is also interesting to do this with the FIDE ratings of the players:

T1, no scrubbing:


$R^2 = 0.96424221670266$

Graph

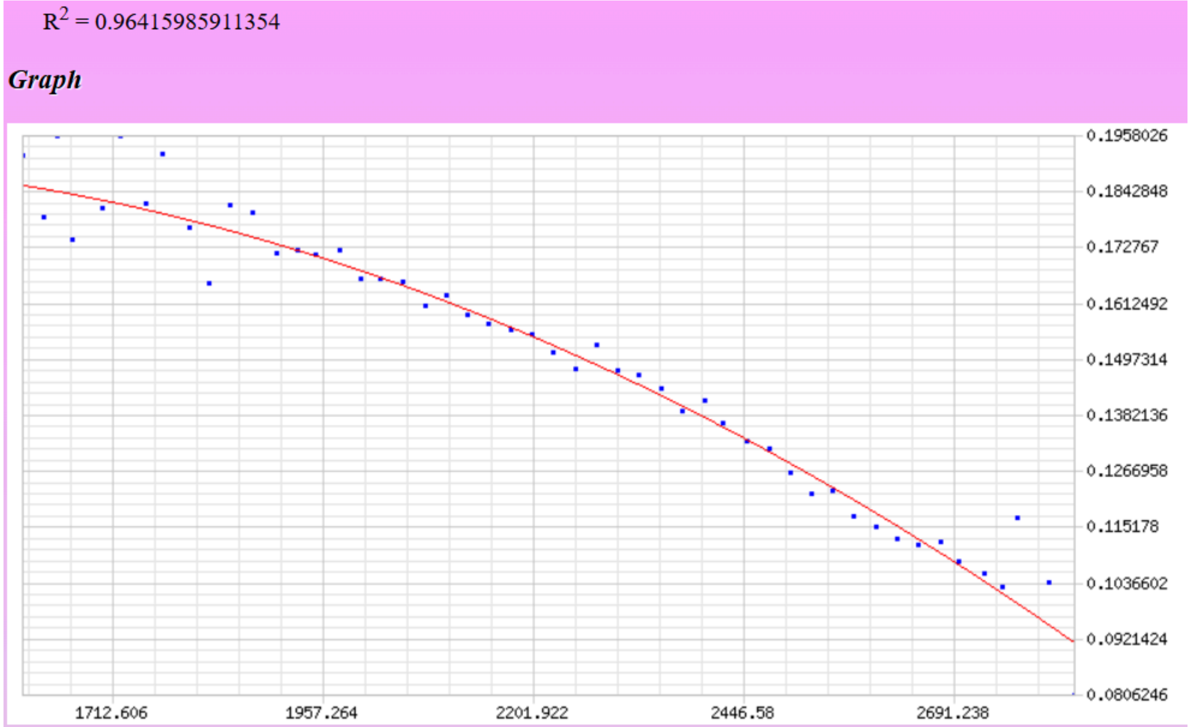
Only 0.012 better than linear in R^2, actually. And when we scrub small buckets below FIDE 1750 and the top bucket (Magnus Carlsen), a line is fine:



$R^2 = 0.96680183618716$

Graph

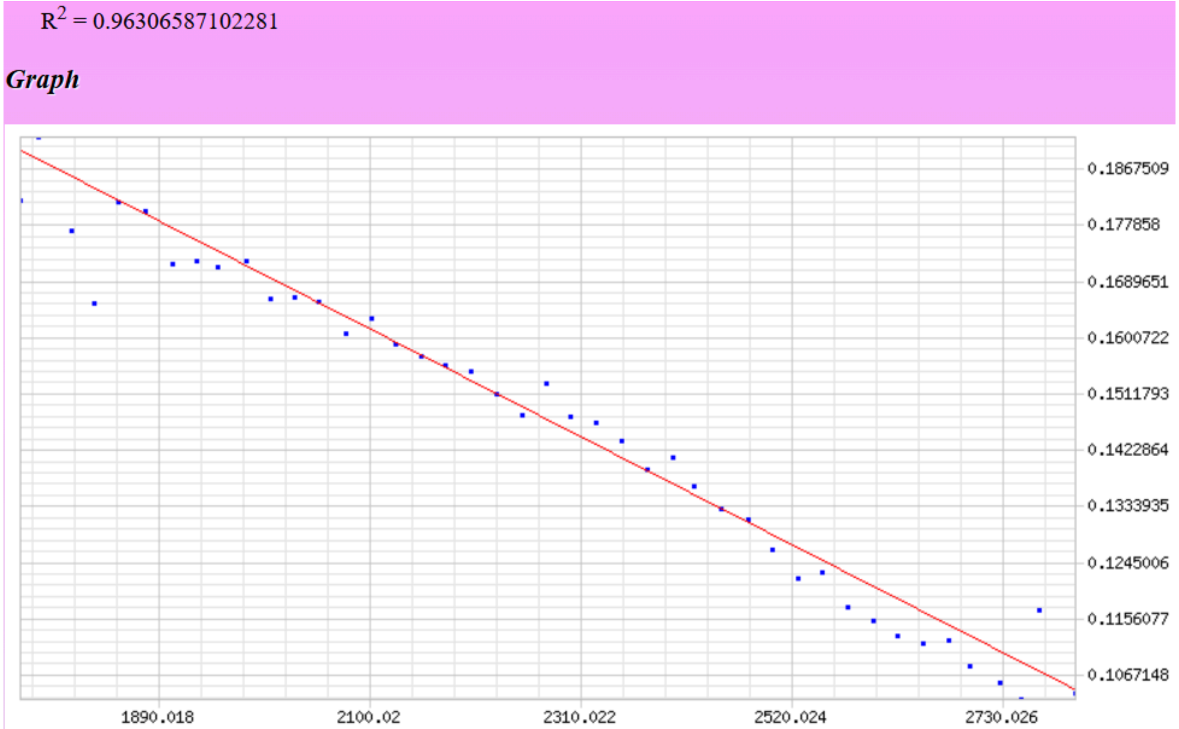
Likewise for ASD, no scrubbing:



$R^2 = 0.96415985911354$

Graph

Looks strongly quadratic---and bending down. But with the same scrubbing, a line is OK again:



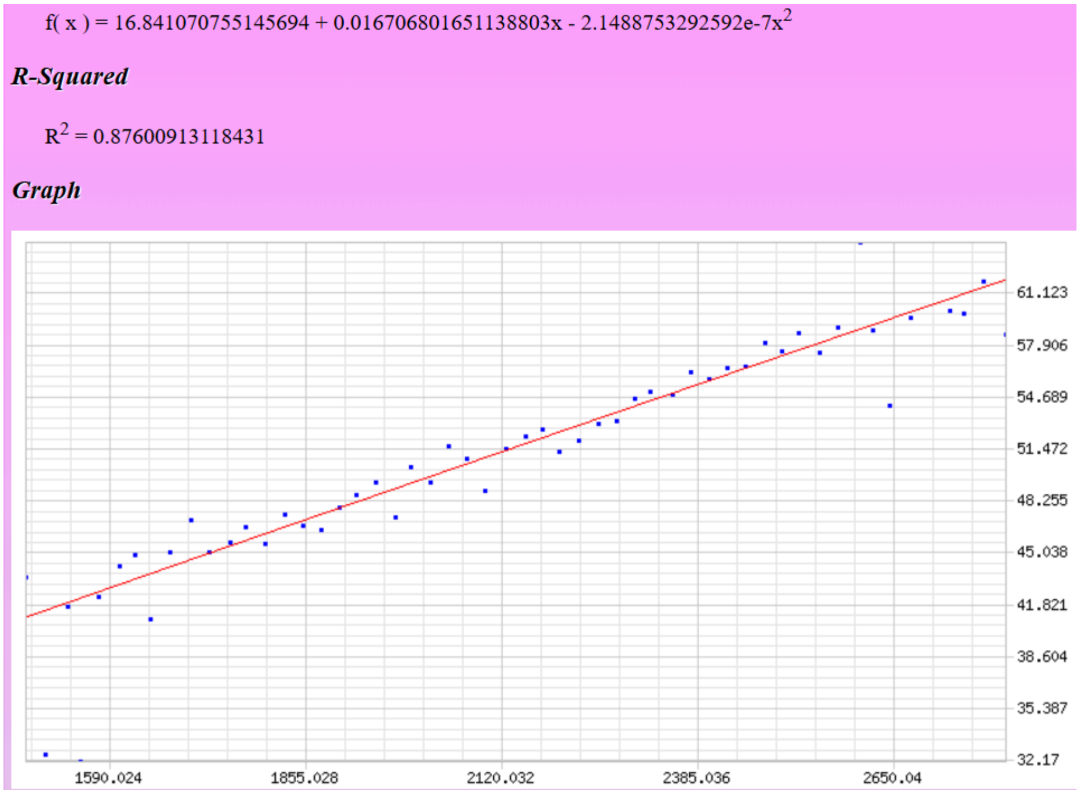
$R^2 = 0.96306587102281$

Graph

Well, this too could also be quadratic. Let's look at the most recent data.

January 2025 Data

Standard OTB tournaments---not including 2024--2025 team events. T1 match, no scrubbing:



f(x) = 16.841070755145694 + 0.016706801651138803x - 2.1488753292592e-7x^2

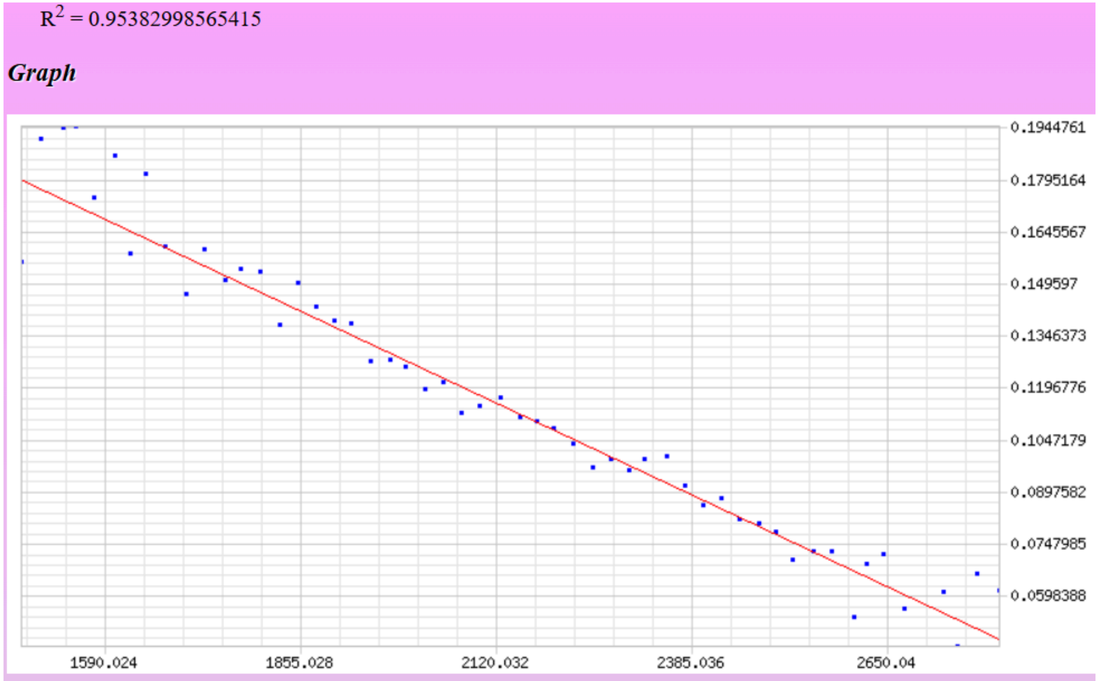
R-Squared

$R^2 = 0.87600913118431$

Graph

A perfectly good line---even though fit shown in quadratic. Scrubbing buckets under 1,000 moves (the top three and below Elo 1600) improves R^2 to 0.920 but does not change the fit itself by much.

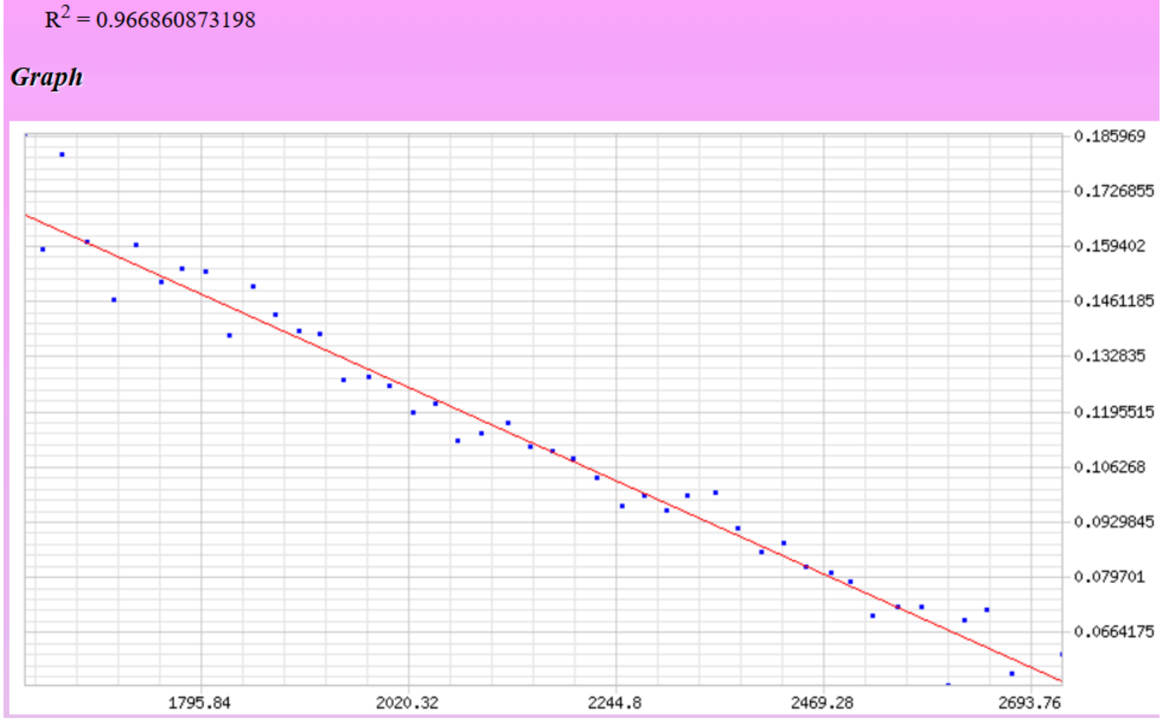
ASD:



$R^2 = 0.95382998565415$

Graph

Even without scrubbing a really good line. A quadratic fit improves R^2 only to 0.960. And with the same scrubbing as above:

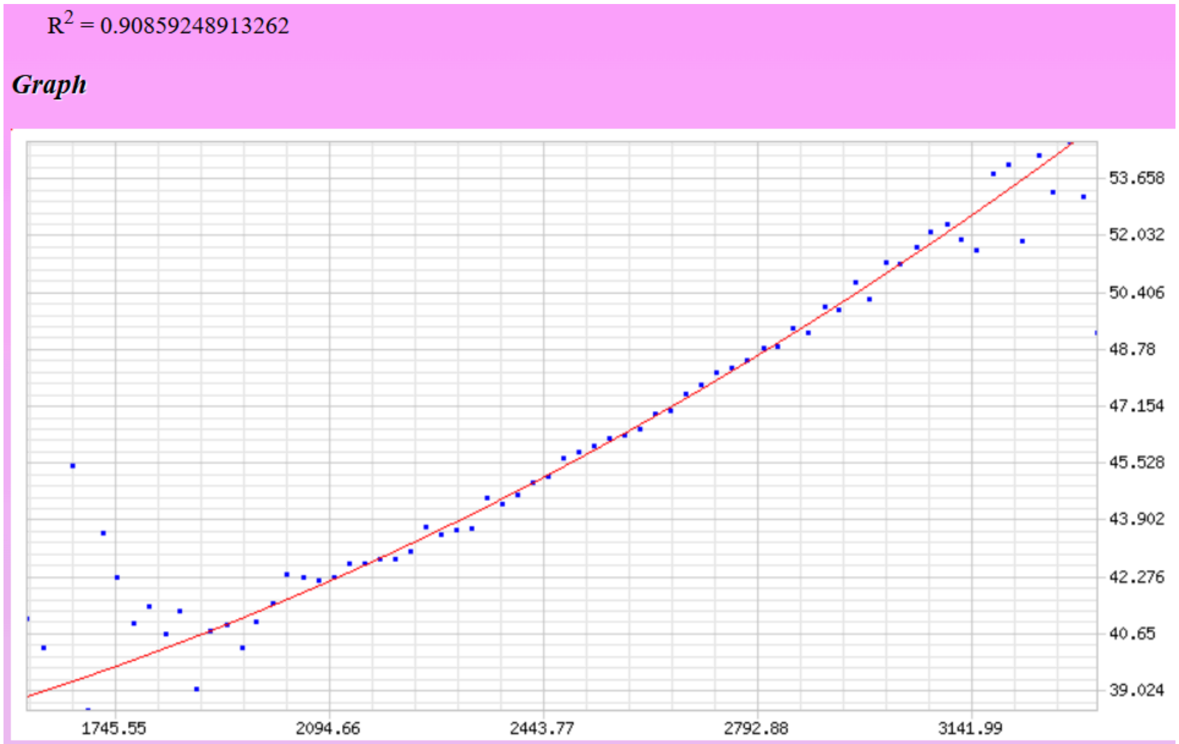


$R^2 = 0.966860873198$

Graph

Now over to Chess.com Titled Tuesday in January 2025, plus 4 Feb. 2025:

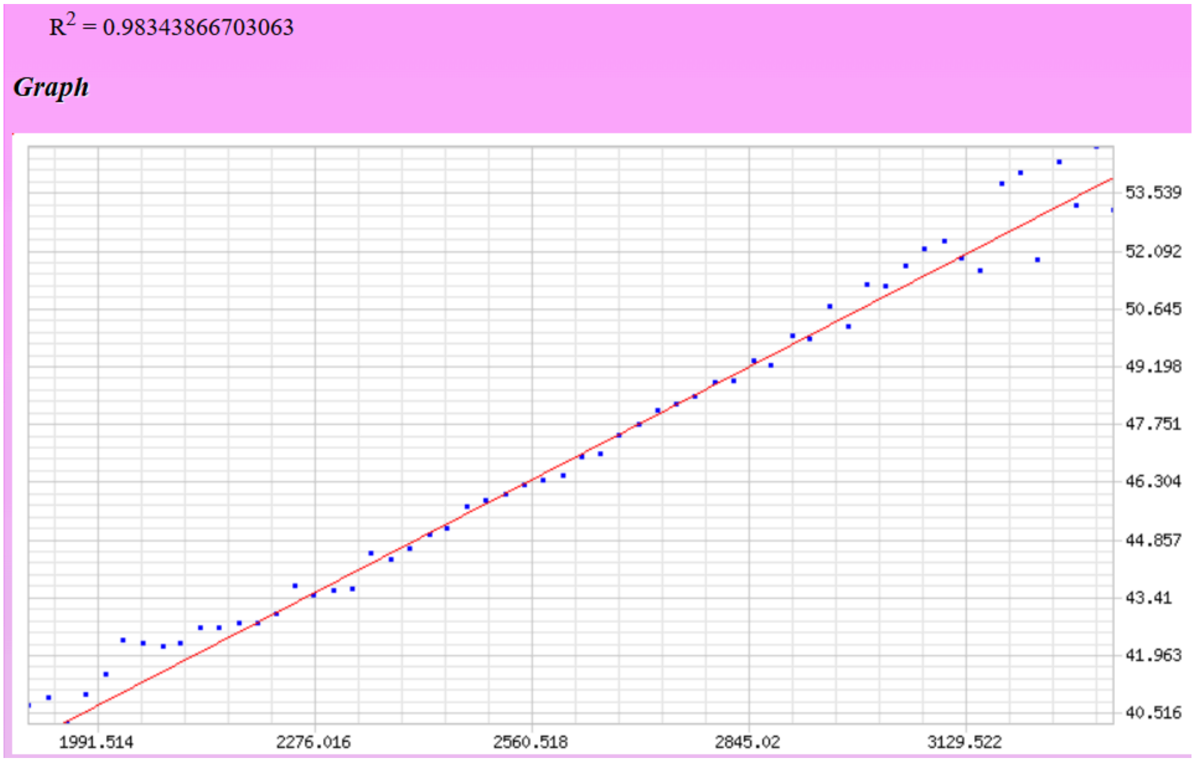
T1-match, no scrubbing, quadratic fit:



$R^2 = 0.90859248913262$

Graph

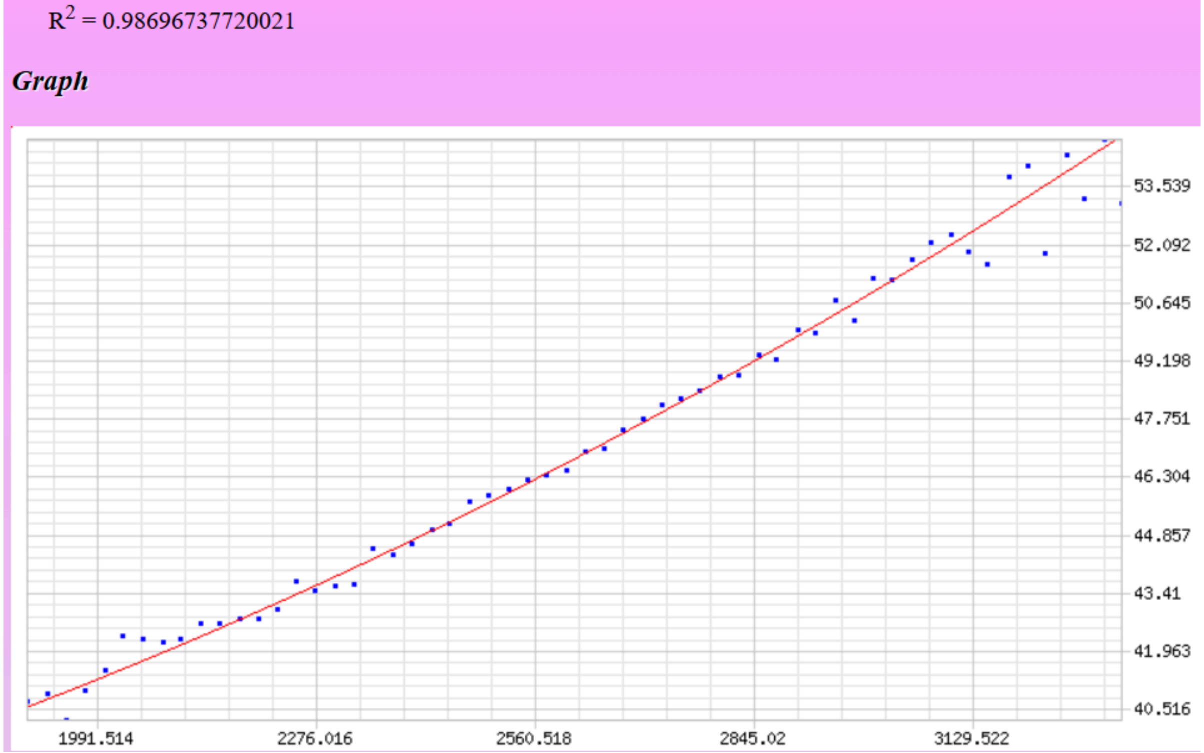
The noise below Chess.com 1900 rating is clear. Scrubbing that and the rightmost bucket with only 715 moves gives



$R^2 = 0.98343866703063$

Graph

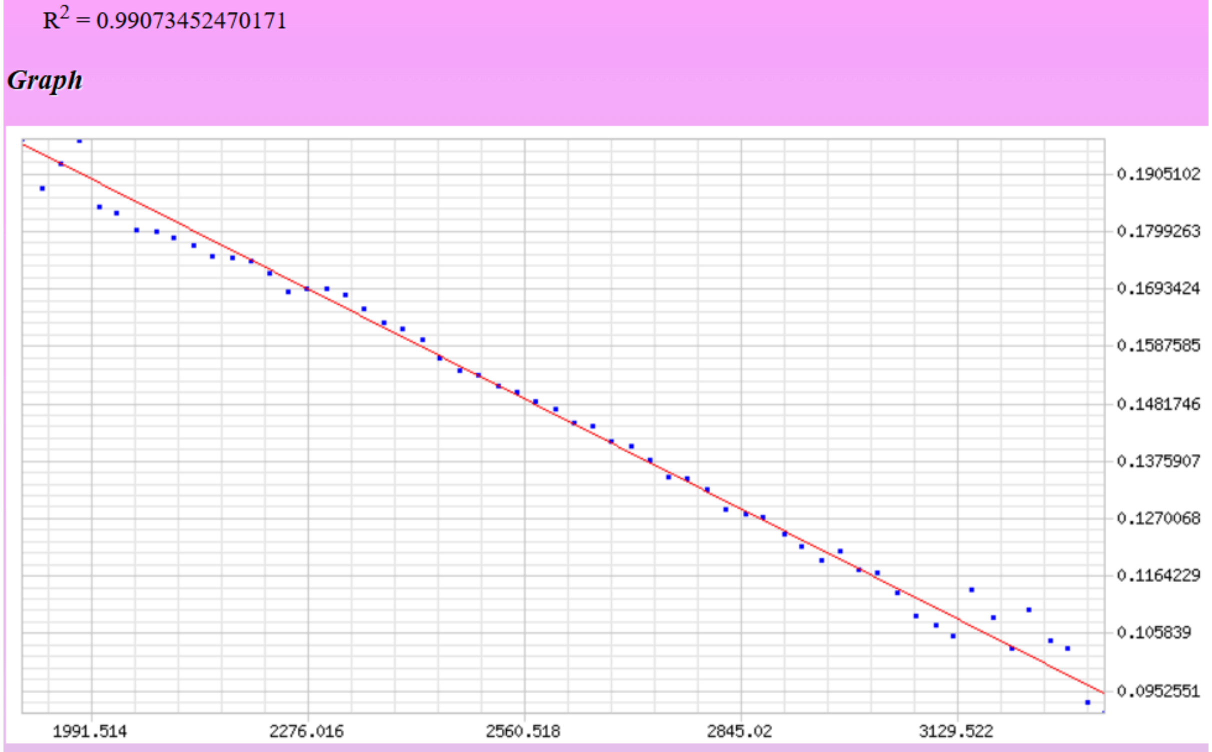
This is fine as a line. The quadratic fit here is only marginally better:



$R^2 = 0.98696737720021$

Graph

And for ASD with Chess.com ratings, same scrubbing:



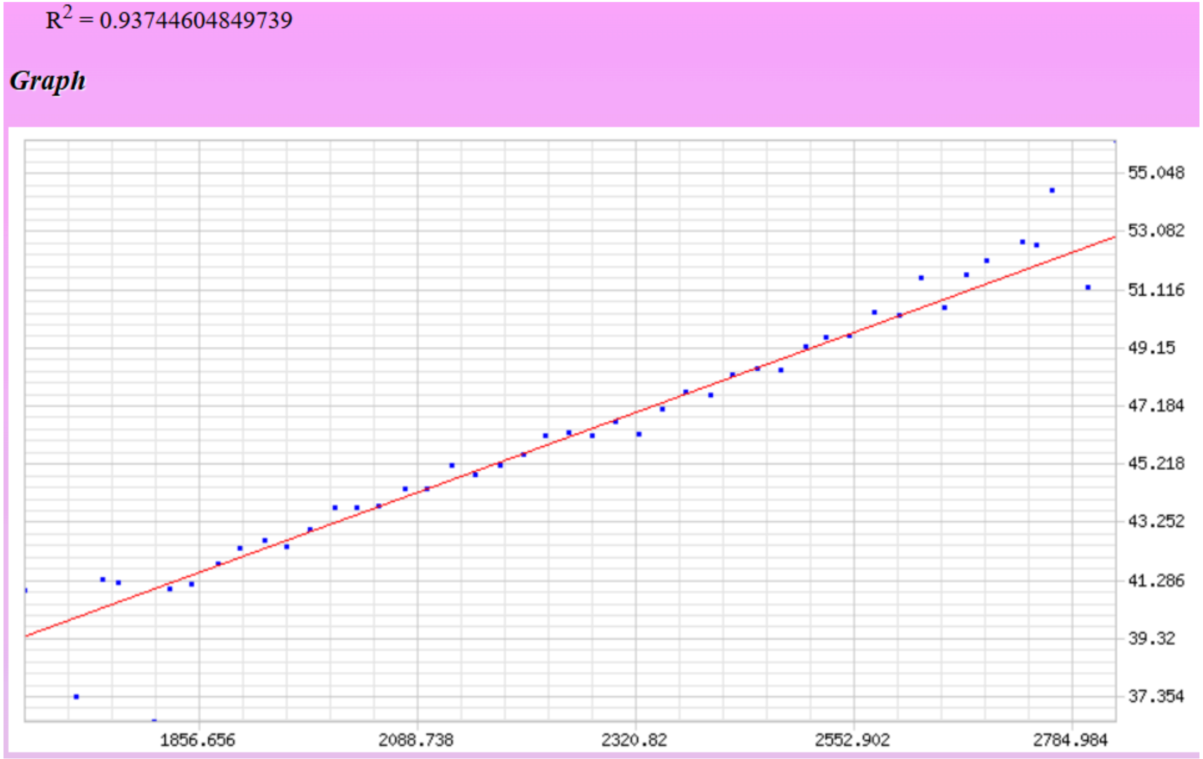
$R^2 = 0.99073452470171$

Graph

A fantastic line. The quadratic fit here actually reports a lower R^2 value.

Last dance is using FIDE Standard ratings for the same players instead (as always, no unrateds):

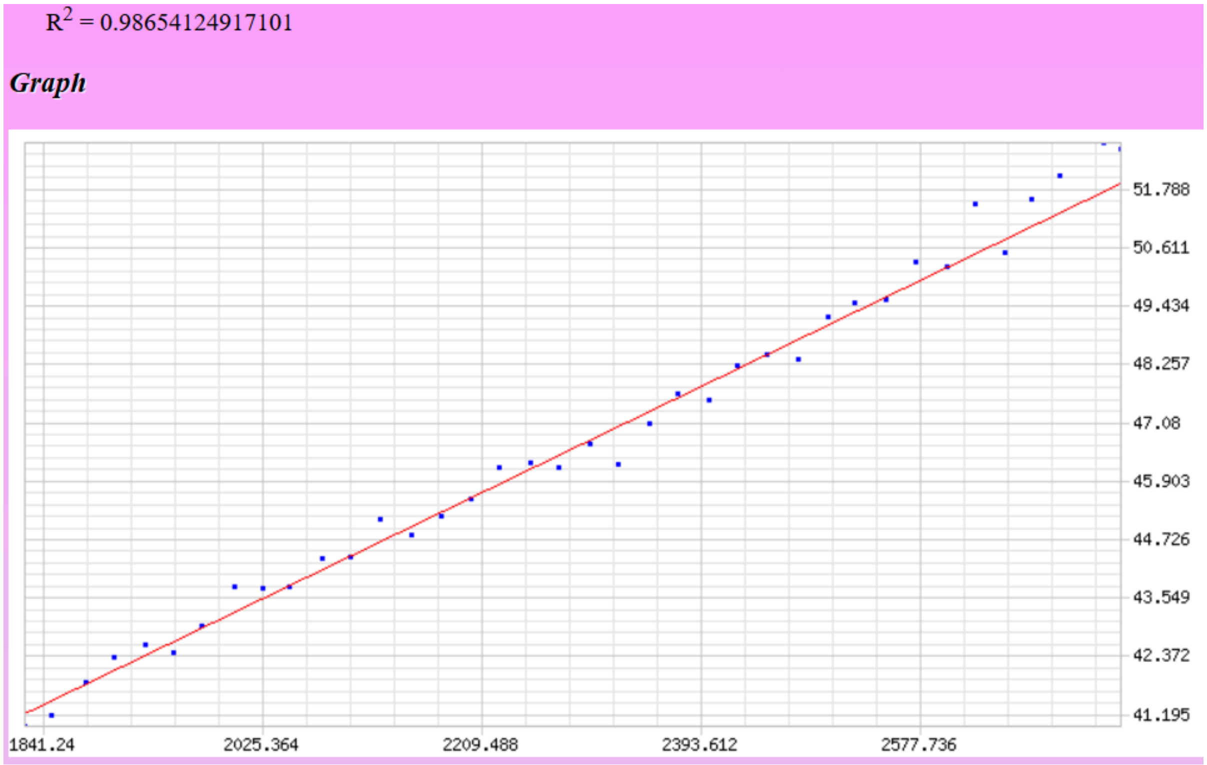
T1-match---no scrubbing, because these FIDE Std. buckets each have at least 4,500 moves except for the second lowest (Elo 1726) with 2,404 moves:



$R^2 = 0.93744604849739$

Graph

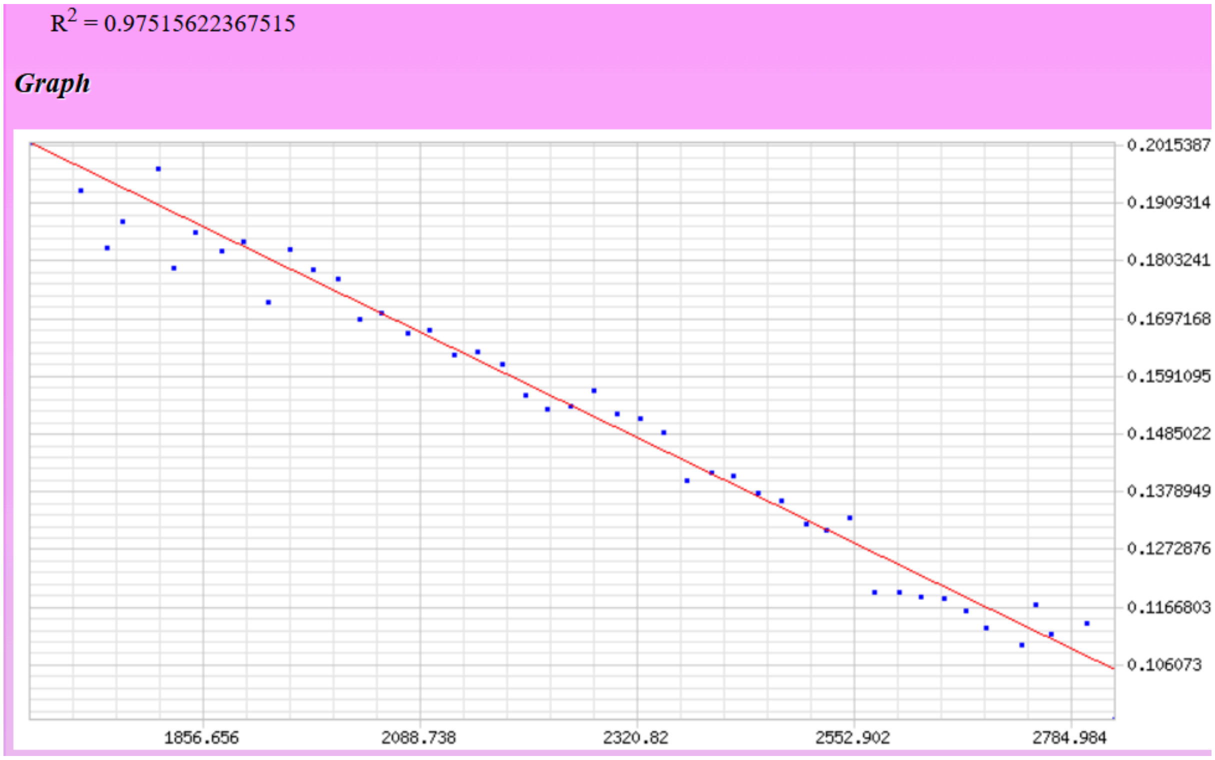
Quadratic fit instead leaves the R^2 basically unchanged. Scrubbing the top three and bottom five leaves all buckets above 23,000 size except for Elo 1825 at just over 10,000 moves:



$R^2 = 0.98654124917101$

Graph

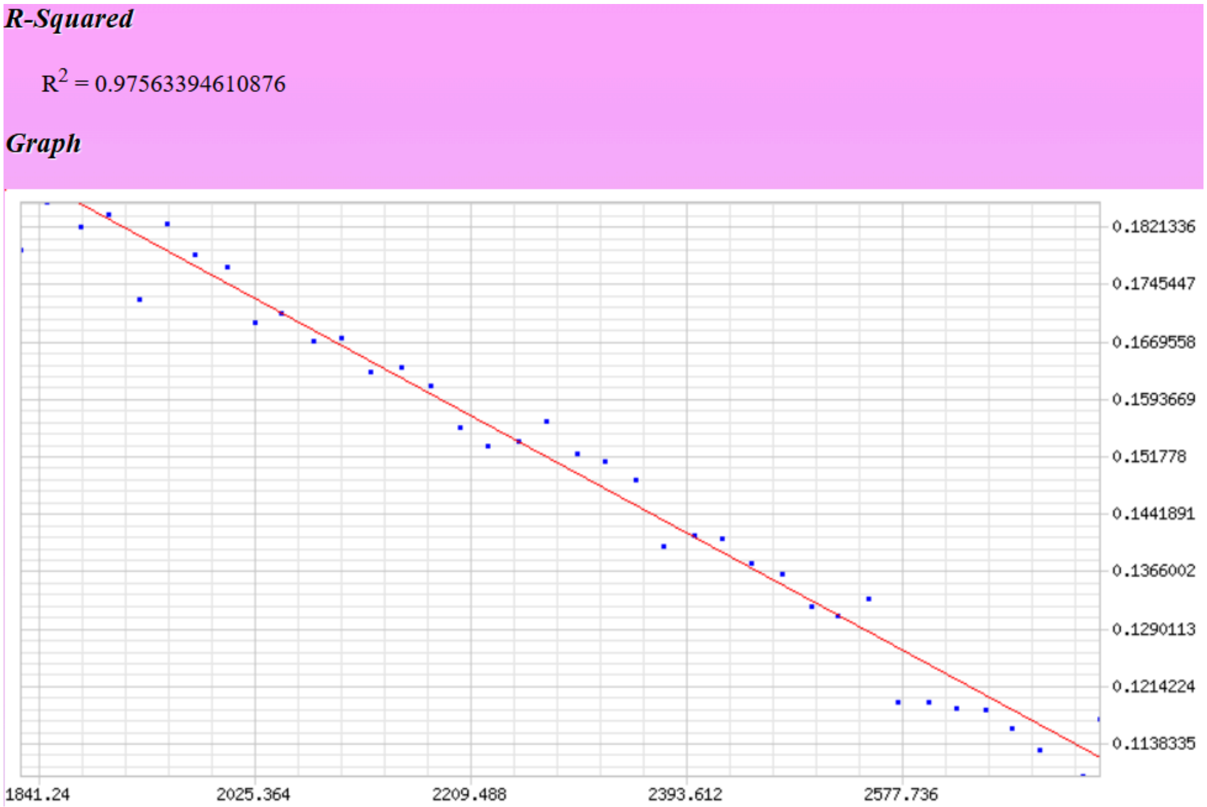
Definitely a line. Finally, for ASD, no scrubbing:



$R^2 = 0.97515622367515$

Graph

Same scrubbing as above:



R-Squared

$R^2 = 0.97563394610876$

Graph

The scrubbing made no change in R^2.

Conclusions

1. Both the T1 and ASD metrics should be strictly linear with Elo rating across the whole spectrum of human skill levels. (The spectrum is limited by the strength of the program at the depth it is operated to. For Stockfish 16 in Single-PV mode, specifying depth at least 20 and at least 5 million nodes, I estimate the resolution as good up to about 3100.)
2. This is true of ASD even though it uses an imperfect logarithmic scaling that is not specific to the engine used. The "purer" ACPL measure---albeit with a cap on the assessed magnitude of any blunder at 4.00---shows highly similar graphs.
3. FIDE Standard ratings were not completely fixed by the Sonas correction at once. Over July---December, the T1-match is linear, but ASD has become curved---possibly a sign of overcorrection. Even in the September Budapest Olympiad and for Oct.-Dec. there is still trace of this curvature. The smaller data from January 2025, however, seems to give strongly linear results for both metrics.
4. The observations between Oct.-Dec. also show up in Titled Tuesday data when FIDE Standard ratings are used, while the Chess.com ratings give linearity. But in January 2025, the FIDE Standard ratings give linearity there as well.

Bottom line: as of end-January 2025, the Sonas correction can be proclaimed in good order. The only lingering niggle is that the ASD line crosses 1400 below 0.20 (in January 2025, about 0.18 in fact), whereas there are FIDE players giving ASD wwll above 0.20. This suggests that 1400 cannot be maintained as the overall rating floor. Ratings should meaningfully be allowed to decline to 1200 if not below.