Regressing Sample Pitching Stats in Order to Estimate True Values

Part I

As I've stated many time before, DIPS ERA, or substituting league average hit rate ($H or (S+D+T)/(PA-BB-SO-HR), for a pitcher's actual sample $H, in order to estimate his true ERA, is a *poor man's regression*. The idea behind DIPS and DIPS ERA is that any deviation in a pitcher's sample $H from some typical $H, is mostly due to luck, as well as the average defense behind the pitcher.

Since it has been shown that pitchers do have *some* control (skill) over their $H, i.e., that less than 100% of those deviations are due to luck and defense, it is not technically correct to substitute some constant $H for a pitcher's sample $H in order to estimate his true ERA. In other words, it is not correct to regress a pitcher's $H 100% towards some constant, as you would if a pitcher had no actual control over his $H. However, because a pitcher does have only a *little* control over his $H, it should be regressed aggressively towards some constant.

Moreover, there is a luck component in a pitcher's other rates as well. Therefore *all* of a pitcher's sample component stats should be regressed towards some constant, and not just $H, in order to estimate his true ERA. The assumption in computing a DIPS ERA and then passing it off as an estimate of a pitcher's true ERA, is that $H can be regressed 100% and that all other components do not have to be regressed at all. This is not really the *assumption* in DIPS of course – the only thing that DIPS *assumes* is that doing it this way yields a *better* estimate of a pitcher's true ERA than either his actual sample ERA or his sample ERC (component ERA).

If we don't take the DIPS shortcut (the *poor man's regression*), and actually apply some regressions to each of a pitcher's sample component stats, we should be able to come up with a much better estimate of a pitcher's true ERA. That should also enable us to come up with a much better projection for a pitcher, since an estimate of a player's true *anything* is essentially the same as his projection.

The important questions are how much do we regress each sample stat, towards what constant should each stat be regressed, and what is the best way to express each stat before doing the regressing? Before I answer those questions, let me add that one of the major shortcomings of DIPS ERA, in terms of estimating a pitcher's true ERA, is that it doesn't consider the sample size of a pitcher's stats.

How much you regress a sample stat depends on (is a function of) two things: One, the element of luck in that stat, which is informed by the distribution of talent in the population vis-à-vis that stat, and two, sample size. Since we know that the element of luck in a pitcher's sample $H is large, we also know that we will regress that stat aggressively when converting a pitcher's stats from "sample to true." That's the whole point of DIPS. Instead of regressing *aggressively*, however, DIPS takes a short cut and regresses *all the way*.

As I said, regression is also a function of sample size, so if we have a large sample of pitcher data, we should regress $H *less* aggressively than if we have a small sample of data, even though much of a pitcher's sample $H is luck. Basically, as with all sample data, the larger the amount of data, the less luck there is, proportionally speaking. Of course, if $H were *all* luck, which it is not, then sample size wouldn't matter. We would simply regress a pitcher's sample $H all the way to some constant (substitute that constant), *a la* DIPS, no matter how large a sample that $H was based on. So DIPS is more correct, at least as far as $H is concerned, for small samples than for large samples.

But what about the other components (BB, SO rate, etc.) in a pitcher's line? I already said that these stats also need to be regressed if we want to estimate a pitcher's *true* stats, and ultimately his true ERA or ERC. Again, DIPS *assumes* that all of a pitcher's stats other than $H are mostly *skill*, such that they don't need to be regressed at all. At least it takes that shortcut. The problem is that, as with $H, the smaller the sample the more these other stats need to be regressed, despite the fact that they have a large skill component. So when you compute a DIPS ERA, the smaller the sample of pitcher data, the more of a mistake you make by not regressing these other stats.

To summarize, the smaller the sample size, the more DIPS is correct in regressing $H all the way, but the less it is correct in not regressing the other stats at all. So no matter what the size of the data, DIPS does not do a very good job of estimating a pitcher's true stats. I suppose that since $H will be regressed fairly aggressively even with large samples of data, but that the size of the regressions for all the other stats varies greatly with the sample size, DIPS works best for large sample sizes. In fact, it doesn't work very well at all for small samples, since it ignores the requisite regressions for every component stat but $H. Of course, if all of a pitcher's other stats (other than $H) , even for a small sample, are around league average, DIPS works fine, as a regression, large or small, will not really change a stat that is near league average in the first place. Then again, if most of a pitcher's stats are around league average, we don't need much help in estimating his true stats anyway.

One thing I was never happy about with DIPS was the arbitrary distinction between hits that stay in the ballpark and home runs. The original purpose of DIPS, I think, was to come up with a *defense independent pitching stat*, hence the name DIPS. However, since we can evaluate defense separately from a pitcher's stats, and adjust those stats accordingly, what we really should be concerned with in terms of regressing a pitcher's stats is *luck* and not defense. In fact, subsequent research on DIPS has suggested that most of the variability in a pitcher's sample $H is *not* due to defense, but to *luck*. I have always advocated changing the name DIPS to LIPS. Unfortunately, it never stuck.

If we focus on luck rather than defense, there is probably a better way to compartmentalize a pitcher's stats that what DIPS does. With DIPS, HR's *per something* (usually BIP) are one category, and non-hr hits *per non-HR BIP* are another. Without going into the details, I have found that a pitcher's *skill* can be better captured by looking at HR's, doubles, and triples as one entity (extra base hits), and singles as another entity. The denominator I use for expressing extra base hits (D, T, and HR) as a rate stat is BIP, and the denominator for singles is BIP minus HR's. Since we are not concerned with defense, the distinction between a home run and a double or triple makes little sense. In fact, it could be argued that the same *skill* is involved in a pitcher allowing a home run as it is in allowing a double or triple (at least a *fly ball* or *line drive* double or triple) – they are all hard hit and/or are long fly balls. The other rate stat I look at (besides the traditional $BB, which is BB/PA and $SO, which is SO/(PA-BB) is HR per extra base hit (HR/(D+T+HR). As it turns out, there appears to be a much larger *skill component* in *extra base hits per BIP* than in *HR per extra base hits*, which does indeed suggest that a pitcher's home runs, doubles, and triples allowed are part of the same skill set, and certainly more related than singles, doubles, and triples, as is the DIPS "grouping".

Let me cut to the chase. I looked at park and defense (using regressed team UZR's) adjusted component pitching stats from 1999 to 2003. From these stats, I looked at all pitchers who had a certain range of TBF's in two years AND had at least 200 TBF's in the following year. The first two years' stats were treated as the *sample stats* and the following year's stats were treated as the *true stats*. The ratio of the two for various sample sizes and for various high and low component rates, as well as the mean of each stat in the *following year*, were used to compute the proper regression coefficients for each component rate stat. In order to *create* or *force* regression, I had to look only at pitchers who had non-average rates for each of the components in the sample years.

Here is how each rate stat is defined:

$BB=BB/PA
$SO=SO/(PA-BB)
$E=(D+T+HR)/(PA-BB-SO)
$HR=HR/(D+T+HR)
$S=S/(PA-BB-SO-D-T-HR)

After *forcing* regressions, here are the interpolated regression values for each of these stats and for various numbers of TBF's:

TBF	**$S**	**$E**	**$HR**	**$BB**	**$SO**
100	.96	.95	.99	.80	.65
400	.85	.75	.95	.50	.25
600	.70	.60	.90	.30	.20
1000	.65	.50	.85	.20	.10
1400	.60	.45	.82	.15	.05
1700	.50	.30	.80	.10	.01

The following are the normalized values towards which each sample rate stat should be regressed, based on the number of sample TBF's. Basically these numbers represent the average quality, vis-à-vis each stat, of the population of pitchers who pitch to X number of batters in any given year. In reality, they are probably more a function of whether a pitcher is a starter or a reliever than how many batters he faced in a season. Therefore, in this chart, TBF is essentially a proxy for starter/reliever status.

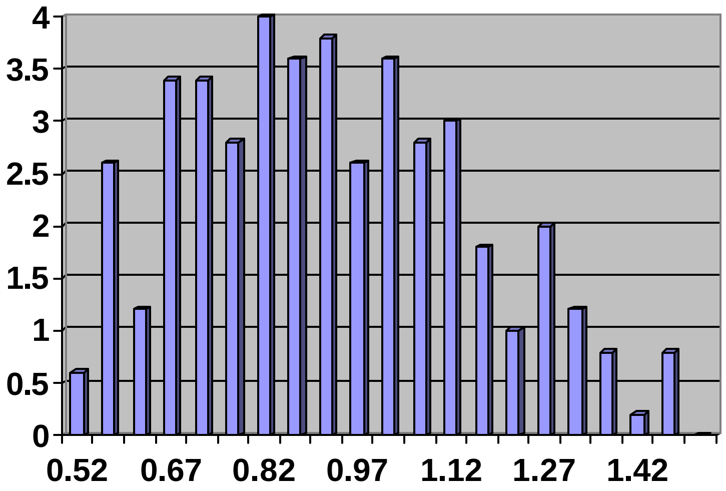
TBF	**$S**	**$E**	**$HR**	**$BB**	**$SO**
100	.98	1.01	1.00	1.09	1.00
400	.97	.96	.97	.99	1.04
600	.99	.96	.95	.95	1.05
1000	1.01	1.01	.98	.90	.95
1400	1.00	.98	.99	.89	.98
1700+	.97	.98	.99	.91	.99

Keep in mind that these are rough estimates based on limited sample data, and that there are some selective sampling problems inherent in the data as well. Nevertheless, I think these are pretty good estimates of the true regression coefficients for these pitcher rate stats.

Also keep in mind that the whole idea of regressing a pitcher's stats toward a particular constant or *mean* is merely a rough approximation of the exact method for estimating a pitcher's true values for each of his stats, and ultimately, his true ERA or ERC. This method of regression (regressing linearly toward a constant) would be pretty close to the *real* method, if the distribution of pitching talent were normal. However, as Bill James likes to point out, the distribution of talent in baseball, and probably in all sports, is not normal, as it is actually a subset of the very tail end of the normal distribution of athletic talent among young males.

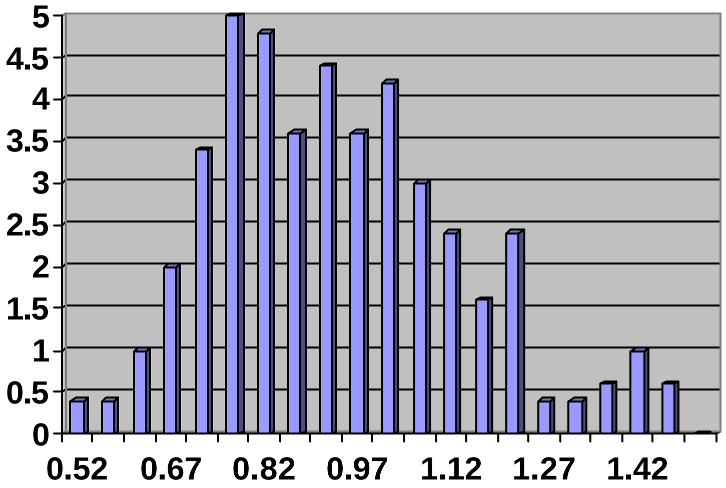
For example, let's look at the distribution in sample $BB (normalized to all pitchers) among all pitchers who had at least 500 TBF's in any year (2000-2002). This group represents most full-time starters. The mean normalized $BB, weighted by the number of TBF's, is .91. If you don't weight by the number of TBF's, the mean is .92. The median is .92. The SD is .078.

Table 1: $BB (Min 500 TBF)



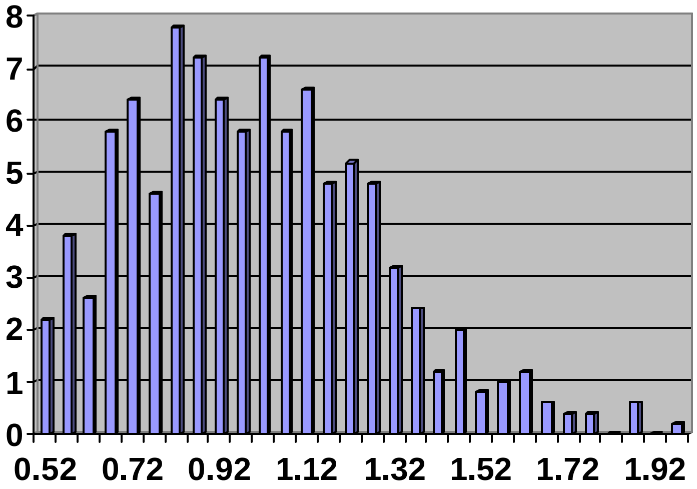
Let's look at the same distribution for $SO, again for only pitchers who had at least 500 TBF's in one season. The mean, weighted by TBF's, is .96. Non-weighed, it is .95. The median is .925. The SD is .068.

Table 2: $SO (min 500 TBF)



If we reduce the minimum number of TBF's to 250, essentially including all regular relievers *and* starters, the $BB distribution looks like this:

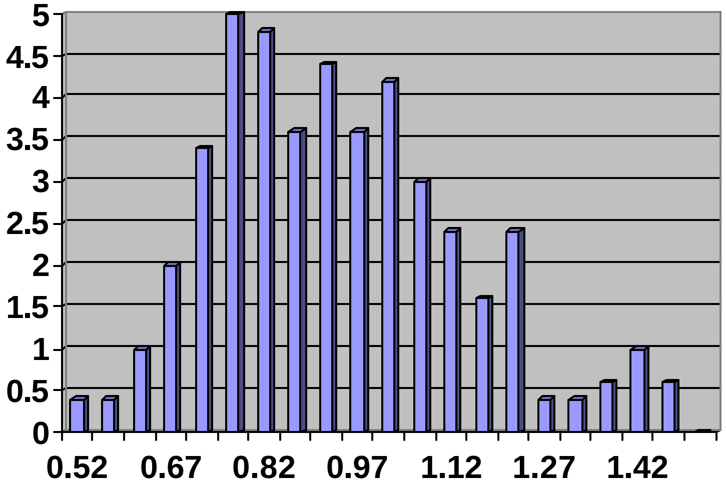
Table 3: $BB (Min 250 TBF)



The weighted mean is .95 and the un-weighted mean is .97. The median is .97 and the SD is .093.

For $SO, it looks like this:

Table 4: $SO (Min 250 TBF)



The weighted mean is .99 and the un-weighted mean is 1.01. The median is .96 and the SD is .094.

Remember, these are distributions of *sample* stats and not *true* stats. In order to see the distribution of *talent* in baseball, we would have to take each player's *sample* stats, convert them into *true* stats by regressing, and then look at the distribution.

Finally, let's do a real life regression in order to convert a pitcher's sample stats into *true* stats and then use those true stats to compute a *true* ERC which will serve as an estimate of that pitcher's true talent or ERA projection:

Let's take one of Kansas City's *young guns*, Jeremy Affeldt. He pitched for KC in 2002 and 2003. We won't consider his minor league performance prior to 2002. Affeldt has a career major league ERA of 4.20. According to baseballreference.com, Affeldt's ERA+ (normalized and park and league adjusted) for those two years is 1.21.

Here are Affeldt's normalized and adjusted components stats for 2002-2003:

TBF	**$S**	**$E**	**$HR**	**$BB**	**$SO**
874	1.09	1.01	.82	1.03	1.19

Interpolating from the charts above, here are the regression numbers and constants for a player with 874 TBF's:

	$S	**$E**	**$HR**	**$BB**	**$SO**
Regression	.67	.54	.87	.24	.14
Constant	1.00	1.00	.97	.92	.97

After doing the regressions, here then are Affeldt's *true* stats:

TBF	**$S**	**$E**	**$HR**	**$BB**	**$SO**
874	1.03	1.005	.95	1.00	1.16

Now let's convert those back into *per 500 PA* (remember, up until now, they all have different denominators), so we can compute an ERC.

TBF	**S**	**E**	**HR**	**BB**	**SO**
874	1.00	.97	.92	1.00	1.16

We can now convert this into a normalized ERC of 1.07 or an ERA+ of 107. Essentially Affeldt's sample ERA+ of 1.21 ended up being regressed to a *true* ERA+ of 1.07, which would also be a good estimate of his 2004 projection, not including any adjustments for experience (pitchers generally get better with experience up to a certain point – so we would project his ERA+ to be slightly better than 107 in 2004).

If you look at the above values and computations, you can see that one *reason* for Affeldt's very good sample ERA+ of 1.21 is his low HR rate *per extra base hit* (.82), relative to his extra base hit rate *per BIP* (1.01). Because my research suggests that a pitcher's extra base hit rate is a much better predictor of his HR rate than is a pitcher's actual HR rate, we would expect his HR rate to regress quite a bit this year, hence the large difference between his sample ERA+ and his *true* or projected ERA+.

The same process can be done to any pitcher's sample stats in order to create an ERA projection. As with batters, a weighted average of a pitcher's previous years' stats should be done first. In other words, Affeldt's 2003 stats should have been weighted more heavily that his 2002 stats. Batter projections involve the same process, although the denominators, and perhaps numerators, for the rate stats should probably be different, triples need to be considered separately from doubles, and the regression numbers and constants will be different of course. In my next article, I will present the proper estimated values for regressing sample batter stats.

Part II

As promised, here are the regression values and constants (towards which you would regress) for batters. The following rate stats are used for batters:

$BB=BB/PA
$SO=SO/(PA-BB)
$HR=HR/(PA-BB-SAO)
$H=(S+D+T)/(PA-BB-SO-HR)
$E=(D+T)/(S+D+T)
$T=T/(D+T)

Batter Regression Values

PA	$H	$E	$T	$HR	$BB	$SO
200	.85	.80	.75	.65	.45	.30
400	.75	.70	.70	.45	.25	.20
600	.70	.60	.60	.30	.15	.15
800	.55	.55	.50	.20	.10	.10
1100+	.50	.50	.40	.10	.10	.05

Batter Regression Constants

PA	$H	$E	$T	$HR	$BB	$SO
200	1.00	1.02	1.18	.92	.88	1.06
400	1.01	1.00	1.04	.81	.94	.98
600	1.01	1.01	.91	.86	.95	.96
800	1.00	1.03	.95	1.05	.96	.95

1100+	1.01	1.03	.92	1.25	1.1	.95

Keep in mind that the regression constants are based on the averages of many batters who for one reason or another had X number of PA's in years one and two (the sample years). As you can see, players who had many PA's in those years tended to be the best overall hitters, with power, not much speed, and a good eye. Batters who had the fewest PA's in the sample years tended to be fast, and with less power and fewer walks – basically the worst hitters.

In practice, when you are regressing a hitter's sample stats in order to do a projection (estimate his true stats), you should use the above regression constants as merely a guide. Remember that these constants are designed to be the *averages of the population from which these batters came.* If, for example, you are regressing the stats of a speedy hitter who also has many PA's (basically a full-time player), you would want to use a $T constant of 1.2 or something like that rather than the .92 or .95 in the chart above for players with lots of PA's in a 2-year period. Just be careful that you don't define a player's *population* from his sample stats. In other words, don't say that your player comes from a population of speedy players (such that you use a 1.2 $T constant) just because he has a very high sample $T. That will completely defeat the purpose of the regression. Be sure that your regression constant is reflected in or defined by some other quality or attribute *other than the stat being regressed*.

The same thing holds true for pitcher regression constants as well. For example, if a particular pitcher throws a 98 MPH fastball, you may want to regress his $SO to a constant other than those indicated by the above charts, regardless of how many TBF's he has, or whether he is a starter or reliever. If you look at the pitcher charts again, you will see that the typical starter with many TBF's will regress towards a $SO of .95 to .98, whereas a typical reliever with fewer TBF's will regress towards a $SO of 1.02 to 1.05. If that 98 MPH pitcher were a starter, you may still want to regress his $SO towards something like 1.05. As with the batters, be careful when *monkeying around* with the pitcher regression constants.

Happy regressing and projecting!