E>xE

Elite above Expectation

A measure of Genetic Merit in Thoroughbred Racehorses

From SW/Rns%, the Average Earnings Index, Standard Starts Index to APEX, measuring the genetic merit of a stallion has been attempted many times, with no metric properly addressing opportunity and thus identifying stallions of true genetic merit.

The below paper, and its proposed metric suggests an answer to the question, "given the opportunity provided, has this sire outperformed expectation as a stallion, and is he a horse of abnormal genetic merit?"

How do we measure expectation?

The mathematics behind a measurement of 'expectation' can be found in the world of professional soccer where Expected Goals (xG) is a popular metric. A description of that measure:

> **Expected goals** (xG) – the number of goals a team or player would be expected to score based on the quality and quantity of shots taken.

Expected goals is a metric which assesses the chance of a shot becoming a goal. It provides a good way to judge the quality of shots taken by a player and a team, since a shot with a 0.4 expected goal (xG) value should be scored 40% of the time. An xG of 1 is the highest value a single shot can be, implying the player has a 100% chance of scoring.

Expected goals as a measure can be beneficial because they increase the sample size used when analyzing soccer. Soccer is a low-scoring game and goals are a rare event. Similarly, in the Thoroughbred world, graded or group stakes winners are a rare event (~2% of the population), and there are many factors that influence their production, so a measure of expectation is a sound approach.

How do we calculate expectation?

In soccer, the generation of xG accounts for the difference in a player shooting from directly in front of the goal (very likely to score) and a player shooting from an acute angle (much less likely to score), distance from goal (further out it is more difficult than close to goal), as well as whether the shot came from a player's head (harder to score) or foot (easier to score). The passage of play that precedes a shot at goal will also have a bearing on the quality of that chance.

Similar to this approach, by looking at historical data we can calculate the average likelihood of each foal by a sire becoming a superior runner by factoring in variables that influence the outcome, that is, under what circumstance was an elite runner produced. The variables that we use to help calculate this 'Expectation of Elite' are:

1. **Generational Interval of Sire** – the difference between the sire's year of birth and the foal's year of birth.
2. **Race Performance of Sire** – the race performance level of the sire.
3. **Foal Rank of Sire** – the foal rank of the foal for the sire. If the first foal, by date of birth, is considered to be noted as '1' and the next born foal by the sire by date of birth is '2', a notation of all foals by the sire with this ranking is possible.

4. **Generational Interval of Dam** – the difference between the dam's year of birth and the foal's year of birth.
5. **Foal Rank of Dam** – the foal number of the foal from the dam.
6. **Race Performance of Dam** – the race performance of the dam.
7. **Production Performance of Dam** – has the dam produced a GSW prior to this foal?

These 7 variables have the greatest influence on the 'opportunity' for an event to occur and effectively builds this into the metric. This is significantly better than the current metric of the 'Comparable Index' which solely looks at the Average Earnings Index for the mares bred to one stallion when bred to other stallions but it doesn't consider that for many stallions they are serving mares that have already produced a stakes winner in their most fertile years (foals 1 through 6) and are now expected to produce a stakes winner from that mare when she is aged. This, in our soccer terms, is like suggesting that a shot on goal taken in front with only the goalkeeper to beat is the same as one taken 30 feet out with a wall of defenders in front. Clearly different opportunity.

Converting the values

The values that we have for many of the data points above will be skewed. For example, every mare has a 1st foal, but not many have a 15th foal. Equally, in order for us to sensibly group the data and allow for enough 'attempts' being made, we need to transform it into a base. To complete this we do three things to each data point, based on a population of data (say the offspring of 500 stallions which if the population of stallions is picked to reflect a commercial population will be somewhere around 100,000+ foals):

1. A Box-Cox transformation. Research is required to find out what Lambda (λ) value is required to created non-skewed data and rescore each value.
2. A Z-Score or Standard Score is created for these new values.
3. This is made into a T-Score where the Z-Score is shifted and scaled to have a mean of 50 and a standard deviation of 10. This then gives each variable a value between 1 and 10.

Obviously, the values of Race Performance of the Sire and Race performance of the Dam are labels such as G1SW, G2SW etc. These can be manually transformed into a rule that creates the same 1 to 10 values (1=unraced to 10=G1SW). The Production Performance of Dam value is a one-hot encoded value with Yes=1 and No=0.

For each foal by each stallion we will then have the following.

	Sire GI	Sire Race Performance	Foal Rank of Sire	Dam GI	Foal Rank of Dam	Dam Race Performance	Dam Production
Foal 1	1	10	5	3	8	1	0
Foal 2	4	8	3	9	5	1	1
Etc...							

With enough data (as suggested, all foals by 500 stallions), we will then be able to 'group' each outcome and calculate the number of times that it has occurred and the number of times that an elite runner has resulted from this group (% Elite to Count). In measuring what is "Elite" it would be horses that are G1, G2 or G3 winners as well as Listed winners and winners that are grade 1 or grade 2 stakes placed. It is also important to have enough data on enough stallions, so we can properly consider each potential outcome.

We then assign this specific value to every foal by the sire. So, in the case of our data above:

	Sire GI	Sire Race Performance	Foal Rank of Sire	Dam GI	Foal Rank of Dam	Dam Race Performance	Dam Produ -ction	xE
Foal 1	1	10	5	3	8	1	0	0.001
Foal 2	4	8	3	9	5	1	1	0.004
Etc...								

Each of the values would represent an expected chance of that outcome being elite. As the industry relies on assortative matings, that is because humans are making decisions on matings similar phenotypes mate with one another more frequently than would be expected under a random mating pattern, the expectation is that there will be particular groups that are significantly more popular (i.e tried more times than normal). Importantly this metric would consider the factors that really influence the potential of the sire to produce a superior runner given the age of the mare, and other variables as discussed above. Once we have an 'expected Elite' or xE value for each foal by a sire, a simple addition of all these values will give us an overall cumulative figure of xE for that stallion (CumxE).

The CumxE figure won't actually represent the value of the stallion. What it will represent is the strength of the mares and the overall opportunity that has been given to the stallion. This figure could be used interestingly to handicap unproven sires.

Calculating Genetic Merit Value (GMV) for a Sire

Each stallion will have three values created based on its stud career:

- CumxE – Cumulative xE – The cumulative score of expected elite (xE) runners
- Elite – the actual number of elite runners sired by the stallion
- E>xE – the number of Elite runners subtracting the expected number of Elite runners (CumxE).

It is important to note that in calculating the E>xE figure, you would only consider data (the CumxE and Elite status) for the foals by a stallion that had 3 or more lifetime starts. This would allow that horse to have an opportunity to perform. To take it back to our soccer analogy, we want to measure when an actual shot on goal has been taken.

To calculate the Genetic Merit Value (GMV) of a sire, we need to do three things to the E>xE values for the population of stallions (500 suggested stallions):

1. A Box-Cox transformation of the E>xE. Research is required to find out what Lambda (λ) value is required to create nonskewed data and rescore each value.
2. A Z-Score or Standard Score is created for these new values.
3. This is made into a T-Score where the Z-Score is shifted and scaled to have a mean of 50 and a standard deviation of 10. This then gives each variable a value between 1 and 10 and thus each stallion a GMV between 1 and 10.

The GMV, which will be a number between 1 and 10, will give a true indication of the genetic merit of a stallion. As only the top 2.2% of all sires will have a score of 8 or above (or greater than 2 standard deviations above the average), this will represent an understandable and mathematically sound way to rate stallion performance.

Calculating Genetic Merit Value for the Broodmare Sire

Given the same mathematics, it is possible to create a GMV for a Broodmare Sire in his role as a Broodmare Sire to assess his true genetic merit in a pedigree page. This is a more simple calculation that revolves around the daughters of the Broodmare Sire, and the outcome of their foals. The data gathered to calculate the merit of a broodmare sire is:

1. **GMV for Broodmare Sire as a Sire** – this will be a value already transformed from 1 to 10.
2. **Generational Interval of Sire** – the difference between the broodmare sire's year of birth and his daughters' year of birth.
3. **Race Performance of his Daughter** – 0=Unraced, 10=G1SW, etc
4. **Race Performance of mares bred to him** – 0=Unraced, 10=G1SW, etc
5. **Generational Interval of Dam** – the difference between the dam of the daughters' year of birth and his daughters' year of birth.
6. **GMV of the Stallion that was bred to his daughter to produce that foal**

	GMV as a Sire	Sire GI	Daughter Race Performance	GrandDam Race Performance	Grand dam GI	GMV of Stallion	xE
Foal 1	7	10	5	3	8	0	0.001
Foal 2	3	8	3	9	5	1	0.004
Etc...							

In exactly the same way that a Stallion GMV is created, with data on enough stallions, the Broodmare Sire GMV can be created that would give a true indication as to the genetic merit of the Broodmare sire as a Broodmare Sire.