Anim. Res. 52 (2003) 79–85

DOI: 10.1051/animres:2003006

Original article

An attempt to predict the earning status of a thoroughbred in France by genealogical data

Bertrand LANGLOIS*, Vincent HERNU

INRA Station de Génétique Quantitative et Appliquée, 78352 Jouy-en-Josas Cedex, France

(Received 12 March 2002; accepted 3 February 2003)

Abstract — The objective of this study was to create indicators to predict which horses had more chances to be placed in races than others. The study was limited to genealogical independent variables (i.e. concerning sires, dams, paternal and maternal half sibs of the horse). It concerned the racing career of 60 851 thoroughbreds born in France between 1980 and 1995 (i.e. annual earnings in flat races of 2 to 5 year-olds and in jumping races of 3 to 5 year-olds). The earning status is the dependent variable; since it is bi-modal, a logistic regression was chosen as the appropriate statistical method. A stepwise procedure with a threshold of 0.001 allows to select adequate independent variables. The out coming figure of this purely phenotypical approach leads to the following conclusions: it is not possible to predict the probability of being placed as a thoroughbred in France only from genealogical data; the most valuable information available to predict this probability, is the previous performance of the horse itself; there is a great dissymetry between the information given by the dams' performances and that given by the sires' performances, which can be ignored; it is also noticeable that the best predictors of a given earning status are those obtained for the family for the same age and discipline. It can be inferred that genetic correlations between age and discipline are different from 1; these correlations are clearly negative between flat and jumping races. This may also be the case in flat races for the correlations between performances at 2 years and performances at 4 and 5 years. These negative relations never appeared in earlier studies since they were based on selected data of placed horses.

horse / thoroughbred / gallop races-earning status / logistic regression

Résumé — **Essai de prédiction du statut gagnant/non gagnant d'un Pur-sang en France à partir de données généalogiques.** L'objectif était de créer des indicateurs pour prédire quels chevaux avaient plus de chance d'être gagnants. L'étude s'est limitée à des variables prédictrices généalogiques (c'est-à-dire concernant le père, la mère et les demi-frères paternels et maternels du cheval). Elle porte sur la carrière de course de 60 851 Pur-sang nés en France entre 1980 et 1995 (c'est-à-dire les gains annuels en plat de 2 à 5 ans et en obstacles de 3 à 5 ans). Le statut gagnant/non gagnant est la variable prédite, comme elle est bi-modale la régression logistique a été choisie comme méthode statistique adaptée. Une procédure pas à pas avec un seuil de 0,001 a permis de choisir les variables

* Correspondence and reprints
Tel.: 33 (0)1 34 65 21 10; fax: 33 (0)1 34 65 22 10; e-mail: Bertrand.Langlois@dga.jouy.inra.fr

prédictrices appropriées. Les grandes lignes qui se dégagent de cette approche purement phénotypique sont les suivantes : il n'est pas possible de prédire correctement si un cheval sera gagnant ou pas uniquement à partir de données généalogiques ; l'information la plus utile pour cela est celle qui peut être obtenue sur le cheval lui-même ; il y a une grande dissymétrie entre l'information apportée par les mères et par les pères, cette dernière pouvant être ignorée ; on peut aussi remarquer que les meilleurs prédicteurs d'un statut gagnant/non gagnant pour un âge et une discipline donnée sont ceux obtenus par la famille pour le même âge et la même discipline. On peut en inférer que les corrélations génétiques entre âge et disciplines sont différentes de 1 ; ces corrélations sont clairement négatives entre courses plates et à obstacles. Cela semble aussi être le cas pour le plat entre statut à 2 ans et statut plus âgé (4 et 5 ans). Ces relations négatives ne sont jamais apparues dans les études antérieures concernant les données sélectionnées des seuls gagnants.

cheval / Pur-sang / statut gagnant-non gagnant / régression logistique

1. INTRODUCTION

The objective of this study was to create indicators to predict which horses had more chances to be placed in races than others. The study was limited to genealogical independent variables, i.e. relative to the horse's family and concerned the racing career in flat races of 2 to 5 year-olds and jumping races of 3 to 5 year-olds.

2. MATERIALS AND METHODS

2.1. Animals

The study concerned 60 851 thoroughbreds born in France between 1980 and 1995. For these horses and their families (sires, dams, paternal and maternal half sibs) the racing earnings were available from 1965 to 2000.

2.2. Description of the main variables

The seven dependent variables are ES, like earning status, followed by the letter F for flat and J for jumping races, and a number indicating age (2; 3; 4; 5). These variables take the value 1 for non placed horses (zero earnings) and 2 for the placed ones.

The independent variables created were numerous. We will just give an overview based on the significant results. The sex is coded 2 for a female and 1 for males and geldings.

The Estes coefficient of success is his annual earning divided by the mean earnings of the horses of the same category (age, sex, year, discipline). An index is then derived by a logarithmic transformation and a standardisation on a mean of 100 and a standard-deviation of 20. This annual index is used to evaluate the level of performance of a horse.

Lev-1: corresponds to the unplaced horses (zero earnings).

Lev-2: corresponds to the poor earnings (index < 100).

Lev-3: corresponds to the moderately good earnings (100 ≤ index ≤ 120).

Lev-4: corresponds to the best earnings (index > 120).

These levels are applied to the family of the horse. The first letters will indicate if the variable refers to the sire (S), the dam (D), the sire of the dam (SD). The following letter indicates the discipline F for flat races and J for jumping races. Then, the coming age (2, 3, 4, 5) and level (see over) are indicated.

SF2 Lev-4 = 1 indicates that the sire of the horse was, at 2 years of age, a very good winner in flat races.

DJ5 Lev-2 = 1 indicates that the dam of the horse was, at 5 years of age, a poor winner in jumping races.

SD F5 L-3 = 1 indicates that the sire of the dam was, at 5 years of age, a moderately good winner in flat races.

The Logarithm of the number of half sibs in each class level (because given the skew distribution of the number of offspring per reproducer, the use of a Log transformation increases the adjustments of the model) is noted as follows:

Np F2-2: is the Logarithm of the number of paternal half sibs (Np) which earned at a low level (-2) in flat races at 2 years of age (F2).

In the same way Nm J5-3 is the Logarithm of the number of maternal half sibs (Nm) which earned at a moderately good level (-3) in jumping races at 5 years of age (J5).

Full sibs are very few and are included as two half sibs.

For a reproducer, the mean of the earning status (ES...) of his offspring gives an equivalent of the percentage of the progeny being placed. For the sire of the dam and performance in flat races with 2 year-olds, we will note this percentage SD.m.F2.

2.3. Statistical method

Since the dependent variable is bimodal, the regression on the categorical data is the appropriate method. The Logistic procedure was chosen using the SAS® [4] software. Significant independent variables were selected using the stepwise option with a threshold of 0.001.

According to the criteria of Akaike and Schwartz, the validity of the models presented was assessed.

They were in all cases better than those considering only the mean.

For the presentation of the results, we chose to explain the effect of each retained independent variable by an odds ratio. These were calculated by taking the exponential of the coefficients found for significant variables in the logistic regression. This is the ratio of the number of favourable cases to the number of non favourable cases. It expresses the chance of a horse to be placed relative to a reference profile (odds ratio = 1).

The determination coefficient of the regression is calculated by a Pseudo-R^2 based on the Log of the likelihood

$D(\mu) = -2\ \mathrm{Log}(\mu) + K_1\ \mathrm{Log}\ I$ for means only

$D(\beta) = -2\ \mathrm{Log}(\beta) + K_n\ \mathrm{Log}\ I$ for the entire model

where K_1, K_n are the number of parameters to be estimated and I is the number of observations. Because $K_1 \neq K_n$ we define only a pseudo – R^2:

$$\text{Pseudo} - R^2 = \frac{D(\mu) - D(\beta)}{D(\mu)}.$$

3. RESULTS

Table I and Table II give the results of the prediction of the earning status of a thoroughbred in France, respectively, for flat and jumping races.

The first observation that can be surprising, is that the influence of the family on the earning status at a given age and discipline has an effect on the horse observed at the same age in the same discipline, sometimes only at a very close age. These influences between ages are absolutely not anarchic.

Secondly, the most important variable in all the models is the previous performance of the horse and particularly during the previous year. When this is not present, for example for 2 year-olds in flat races (pseudo $R^2 = 0.13$) or for 3 year-olds in jumping races (pseudo $R^2 = 0.12$) the prediction is bad. The percentage concordant is less (nearly all horses are predicted as non placed, and the percentage concordant is very close to the percentage of non placed horses). The percentage of false positives (horses predicted as placed, being not placed) is very high (near 0.50) and the percent of the Sommer's D, Gamma,

Table I. Variables and corresponding odds ratio retained by a logistic regression on the stepwise mode with threshold 0.001 to predict the earning status in flat races.

Dependent variable	ESF2	Odds ratio	ESF3	Odds ratio	ESF4	Odds ratio	ESF5	Odds ratio
Independent	male	1	ESF2	7.537	ESF3	11.98	ESF4	17.308
variables	female	1.083			ESJ3	0.438	ESF3	3.836
	SJ5 Lev1	1	Male	1	ESF2	1.556	ESJ4	0.661
	SJ5 Lev2	0.431	female	0.844				
	SJ5 Lev3	0.497	Np. F3-3	1.305	male	1	male	1
	SJ5 Lev4	0.496	Np. F3-4	1.099	female	0.733	female	0.564
					SJ4 Lev1	1	SJ3 Lev 1	1
	DJ4 Lev1	1	Np. J3-2	0.805	SJ4 Lev4	0.534	SJ3 Lev2	0.229
	DJ4 Lev2	0.848						
	DJ4 Lev3	0.833	Nm. F3-2	1.300	DF4 Lev1	1	DF4 Lev1	1
	DJ4 Lev4	0.677	Nm. F3-3	1.492	DF4 Lev3	1.197	DF4 Lev2	1.241
			Nm. F3-4	1.546	DF4 Lev4	1.384	DF4 Lev3	1.182
	DF2 Lev1	1					DF4 Lev4	1.376
	DF2 Lev2	1.175	Nm. J3-2	0.883	SD F2 L-1	1		
	DF2 Lev3	1.344			SD F2 L-4	1.134	Np. F2-3	0.840
	DF2 Lev4	1.408	SJ3 Lev1	1			Np. F4-3	1.245
			SJ3 Lev2	0.619	Np. F5-3	1.062		
	SD.m.J5	0.417	SJ3 Lev4	0.676	Np. F5-4	1.056	Nm. F5-2	1.193
	SD.m.F2	1.599					Nm. F5-3	1.198
			DF3 Lev1	1	Nm. F5-2	1.226	Nm. F5-4	1.375
	Np. F2-2	1.228	DF3 Lev2	1.054	Nm. F5-3	1.351		
	Np. F2-3	1.360	DF3 Lev3	1.168	Nm. F5-4	1.312		
	Np. F2-4	1.209	DF3 Lev4	1.424				
					Nm. J3-3	0.914		
	Np. J5-2	0.813	SD F3 L-1	1				
	Np. J5-3	0.938	SD F3 L-2	0.807				
			SD F3 L-3	0.846				
	Np. F4-2	0.903						
	Np. F4-4	0.896	SD m. F3	1.588				
	Nm. F2-2	1.578						
	Nm. F2-3	1.695						
	Nm. F2-4	1.651						
	Nm. J5-3	0.898						
	Nm. J5-4	0.810						
Pseudo- R^2	0.13		0.30		0.35		0.46	

ES: earning status; F2, F3, F4 and F5: flat races at 2, 3, 4 and 5 years of age, respectively; J2, J3, J4 and J5: jumping races at 2, 3, 4 and 5 years of age, respectively; S, D and SD: variable refers, respectively, to sire, dam and sire of the dam; Lev1: zero earning; Lev2 (or -2): poor earnings; Lev3 (or -3): good earnings; Lev4 (or -4): the best earnings; Np: number of paternal half sibs; Nm: number of maternal half sibs.

Table II. Variables and corresponding odds ratio retained by a logistic regression on the stepwise mode with threshold 0.001 to predict the earning status in jumping races.

Dependent variable	ESJ3	Odds ratio	ESJ4	Odds ratio	ESJ5	Odds ratio
Independent variables	ESF2	2.224	ESJ3	10.972	ESJ4	12.622
			ESF3	2.245	ESJ3	2.231
	male	1			ESF4	2.491
	female	0.742	male	1		
			female	0.602	male	1
	DJ4 Lev1	1			female	0.552
	DJ4 Lev2	1.206	DJ4 Lev1	1		
	DJ4 Lev3	1.495	DJ4 Lev2	1.240	DJ5 Lev1	1
	DJ4 Lev4	1.832	DJ4 Lev3	1.444	DJ5 Lev2	1.318
			DJ4 Lev4	1.406	DJ5 Lev4	1.523
	SD. m. J4	2.570				
			SD m. J5	3.759	SD m. J5	5.745
	Np. J3-2	1.157				
	Np. J3-3	1.354	Np. J5-2	1.174	SD F3-1	1
	Np. J3-4	1.232	Np. J5-3	1.236	SD F3-2	1.310
			Np. J5-4	1.157		
	Np. F4-2	0.874			Np. F2-2	0.904
	Np. F4-3	0.929	Np. F2-2	0.894	Np. F2-3	0.812
	Np. F4-4	0.806	Np. F2-3	0.881	Np. F2-4	0.926
			Np. F2-4	0.896		
	Nm. J3-2	1.648			Np. J4-2	1.214
	Nm. J3-3	1.947	Nm. J4-2	1.473	Np. J4-3	1.223
	Nm. J3-4	2.391	Nm. J4-3	1.710	Np. J4-4	1.059
			Nm. J4-4	1.718		
	Nm. F4-3	0.872			Nm. F3-3	0.873
	Nm. F4-4	0.807	Nm. F3-3	0.927	Nm. F3-4	0.880
			Nm. F3-4	0.779		
					Nm. J5-2	1.581
					Nm. J5-3	1.442
					Nm. J5-4	1.351
Pseudo- R^2	0.12		0.30		0.39	

For abbreviations, see Table I.

Tau – a and c using the results of the number of concordant and discordant pairs are smaller.

For the other ages, pseudo – R^2 increases notably: 0.30; 0.32; 0.46 respectively for flat races from 3 to 5 years of age 0.30, and 0.39 for jumping races at 4 and 5 years of age. This increase resulted, however, mainly in a diminution of the percentage of placed horses. The prediction was more accurate because not placed is mainly predicted. The percentage of false positives is at the minimum (26%) for 3 year-olds in flat races. For the other situations in predicting a placed horse, a false diagnostic is made around 4 times out of 10. This is indeed better than 1 time out of 2 when previous performances are not taken into account, but it remains rather low.

Concerning the sex, females are more disadvantaged as they advance in age. With 2, running separately from the male, they

benefit from a slight advantage odds ratio (o.r. = 1.083) which disappears with 3 year-olds (o.r. = 0.844) and decreases regularly with 4 (o.r. = 0.733) and 5 year-olds (o.r. = 0.564). So at this age, we can consider all other things equal that a male has 1/0.56 = 1.78 more chances to be placed than a female. This phenomenon is even greater in jumping races o.r. = 0.742; 0.602; 0.552, respectively for 3, 4, and 5 year-olds.

Concerning the performance of the parents, one can notice that the probability of being placed increases with the level of dam performance at the same age in the same discipline. When this increase is not observed, the coefficients are not significantly different. As an example let us see the effect of the dam level in flat races of 3 year-olds on the probability of being placed at the same age and discipline for the offspring.

The odds ratios are 1; 1.054; 1.168; 1.424, respectively, for DF3 lev-1; -2; -3; -4. To have an unplaced dam is the modality of reference (o.r. = 1). So we can see that the higher the performances of the mare, the higher are the chances of her offspring to be placed.

For the equivalent performances of the sire, it is a surprise, they never appear to be significant. The only cases are the negative effects of jumping race performances on the probability of their offspring to being placed in flat races. The performances of the sires therefore does not seem to be informative. For foreign stallions, all information is also not always available. A good performer could therefore appear as not being placed.

This therefore does not mean that the stallions had no genetic influence as shown by the more balanced figure given by the levels of paternal (Np) and maternal (Nm) half sibs. The production of sires are better indicators of their genetic differences than their own performances (at least in France).

It is also remarkable that the level of the sire of the dam mostly plays a significant role. However his effect is sometimes paradoxical. For example, for flat races for 3 year-olds, the performances of the sire of the dam have a significant negative effect; in parallel the dams sires percentage of placed offspring (m) has a significant positive effect.

For the number of half sibs if multiplied by e ($\simeq 2.7$), the Logarithm will increase by 1 and the chance of being placed will be multiplied by the odds ratio: for example, multiplied by 1.305 for 3 year-olds in flat races for a number of 3 year-old paternal half sibs having a moderately good level (lev-3) in flat races moving from 10 to 27 or from 3 to 7 approximately.

One can also notice that there are clear negative genetic relationships between flat and jumping racing performances. The only exception is the positive relationship between the performance of the sire of the dam in flat races at 3 years (SDF3-2) with E5J5 (o.r. = 1.310).

Finally for two year-olds, an opposition is sometimes found with the performances of old horses, as with the following: the negative effect on ESF2 of the number of paternal half sibs placed in flat races at 4 years of age (o.r. = 0.903 and 0.896 for NpF4-2 and NpF4-3); or with the negative effect of the number of paternal half sibs placed in flat races with 2 year-olds (NpF2-3; o.r. = 0.840) on the probability to be placed in flat races with 5 year-olds (ESF5).

4. CONCLUSION

This statistical study is the first study based on all born horses instead of placed ones.

It concerns more than 60 000 born horses with their whole career in flat and jumping races in France. The degree of accuracy is therefore very good.

The outcoming figure of this purely phenotypical approach leads to the following conclusions:

– it is not possible in France to predict the probability of a thoroughbred of being placed only by genealogical information. This is not very new for geneticists but may lead horse buyers to more prudence in giving the price for a yearling;

– the most valuable information that can be obtained to predict the probability of being placed is the previous performances of the horse, i.e. the information you can get on the horse itself. In this case you make better predictions (rather good in general) because your prognostic is mainly not placed and the percentage of placed horses is low. But if you predict placed, you are right only 6 times out of 10 which is not very good...;

– there is a great dissymetry between the information given by the dam's performances which is in all cases very significant and that given by the sire's performances which can be totally ignored. This phenomenon is probably due to the selection of stallions and the lack of information due to international exchanges and is not taken into account when modelling roughly with an animal model;

– the dissymetry disappears when you consider the genetic effect at the offspring level. A certain genetic background of the probability to be placed can be assumed;

– it is also noticeable that the best genetic predictors are those obtained for the family for the same age and discipline. Genetic correlations between age and discipline therefore seem significantly different from 1;

– the figure given for these correlations is clearly negative between flat and jumping races, and may also be in flat races between performances of 2 year-olds with performances of 4 and 5 year-olds.

Theses negative relationships have never appeared before in studies based on selected data based on placed-horses [1, 2, 3].

REFERENCES

[1] Cunningham E.P., Genetics of performance traits – Thoroughbred, in: Bowling A.T., Ruvinsky A. (Eds.), The Genetics of the horse, Cabi. Publishing, Chap.15, 2000, pp. 411–418.

[2] Langlois B., Heritability of racing ability in thoroughbreds – a review, Livest. Prod. Sci. 7 (1980) 591–605.

[3] Langlois B., A consideration of the genetic aspects of some current practices in thoroughbred horse breeding, Ann. Zootech. 45 (1996) 51–51.

[4] SAS Institute Inc., SAS/STAT® User's Guide, Version 8, SAS Institute Inc., Cary, NC, Chap. 39, 1999, pp. 1903–2042.