Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2012:2**,
http://www.math.su.se/matstat

Statistisk analys av rekord i löpning

Hilja Brorsson*

Februari 2012

Sammanfattning

Detta arbete syftar till att genomföra en statistisk analys av en uppsättning rekord i löpning för att undersöka hur olika faktorer inverkar på tiden. Fem faktorer väljs ut att ingå i studien: distans, rekordtyp, kön, åldersgrupp och arenatyp. Inledande undersökningar visar att det innan analysen är nödvändigt att korrigera datamaterialet genom tillämpning av log-transformation och justering för observerade effekter av reaktions- och accelerationstid. För att finna en modell som i största möjliga mån fastställer variationsorsakerna i tiden görs därefter jämförelser av flertalet olika modellformuleringar genom tillämpning av metoder inom varians- och kovariansanalys. Distans visar sig direkt vara den variabel som har den överlägset största inverkan på tiden, vilket stämmer överens med vad som inledningsvis förutspås. Till följd av detta fokuserar den huvudsakliga analysen på att undersöka vilka variabler som bidrar till att förklara den mindre delen återstående oförklarad variation och hur en modell lämpligen bör formuleras för att i största möjliga mån beskriva sambandet. Inom ramen för linjär regression resonerar vi oss fram till en fjärdegradsmodell där sammanlagt åtta förklaringsvariabler ingår. Genom att betrakta distans som en kategorivariabel fås även en alternativ modell med 14 förklaringsvariabler, varav 10 av dessa representerar uppsättningen av distanser. I båda modellerna beskrivs responsvariabeln av samma uppsättning faktorer. Vidare observeras det att de fyra kategoriska faktorernas effekter i logskala fås i hög grad oberoende av distans. Den genomförda variansanalysen avslutas med en diskussion kring de två föreslagna modellernas lämplighet genom tillämpning av olika metoder för modellkritik, varefter en sammanfattning görs av de slutsatser som kan dras beträffande faktorernas effekter. Arbetet avslutas med en diskussion kring genomförda analyser och resultat, där även eventuella brister och begränsningar i utförandet berörs.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: hilja_@hotmail.com. Handledare: Rolf Sundberg.

Abstract

The aim of this report is to perform a statistical analysis on a set of running records in order to examine how different factors affect the running time. Five factors are included in the study: distance, type of record, sex, age group and type of arena. Initial examination shows that it's necessary to correct the existing data before the analysis by application of log transformation and adjustment of observed effects of reaction and acceleration time. To find a model which can establish as much causes of variation in the time as possible, comparisons are made of several different model formulations by application of methods in variance and covariance analysis. Distance immediately shows to be the variable which by far has the greatest effect on the time, which corresponds well with what is intuitively predicted. As a result of this, the continued analysis primarily focuses on examining which variables contributes to the explanation of the small amount remaining unexplained variation and how an adequate model should be formulated to describe the relationship to the greatest extent possible. Within the framework of linear regression, we arrive at a 4^{th} degree model containing a total of eight explanatory variables. By regarding distance as a categorical variable, an alternative model is obtained as well, with 14 explanatory variables, of which 10 represents the set of distances. In both models, the response variable is described by the same set of factors. In addition, the log-scale effects of the four categorical factors were seen to be essentially independent of distance. The performed analysis of variance is ended with a discussion on the suitability of the two proposed models by application of various methods of model criticism, after which we summarize the conclusions that can be drawn concerning the factors effects. The report ends with a discussion on the performed analysis and results, where possible inadequacies and limitations in the performance also are mentioned.