Stockholms universitet

Statistisk analys av rekord i löpning

Hilja Brorsson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2012:2
Matematisk statistik
Mars 2012

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

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.

Förord och tack

Detta examensarbete omfattar 15 högskolepoäng och leder till en kandidatexamen i biomatematik med inriktning mot matematisk statistik vid Matematiska institutionen på Stockholms universitet.

Jag vill rikta ett stort tack till min handledare Professor Rolf Sundberg vid Matematiska instutionen på Stockholms universitet för all hjälp och vägledning under arbetets gång.

Stockholm, mars 2012
Hilja Brorsson

Innehållsförteckning

1 Introduktion

1.1 Inledning

Löpning på slät bana är en av de mest grundläggande och populära idrottsgrenarna inom friidrotten. Aktuella nationella och internationella rekord förbättras fortlöpande och inte sällan uppdateras de med endast någon eller några enstaka hundradelar. Varje löpare har sin egna personliga stil och fysiska förutsättningar, men det finns även en uppsättning gemensamma faktorer som inverkar på prestationerna. Det är lätt att förutspå att distansen rimligtvis bör har den överlägset största betydelsen för rekordet, men vilka andra faktorer har effekt på en rekordinnehavares prestation och hur kan detta samband beskrivas?

1.2 Bakgrund

Officiella världsrekord i löpning ratificeras av den internationella organisationen för friidrottsförbund, International Association of Athletics Federations (IAAF). IAAF grundades i Sverige 1912 och organiserar idag många stora mästerskap, bl.a. friidrotts-VM (hämtat från http://en.wikipedia.org/wiki/International_Association_of_Athletics_Federations). Officiella svenska rekord ratificeras av Svenska Friidrottsförbundet (SFIF) som är medlemsorganisation under IAAF. De gällande reglerna för friidrottstävlingar fastställs av IAAF och för att resultat och eventuella rekord ska erkännas officiellt måste dessa följas av de 212 medlemsländerna. Utöver IAAF:s tävlingsregler finns även vissa nationella bestämmelser, vilka främst gäller för barn- och ungdomstävlingar (hämtat från http://www.friidrott.se/alltom/regler/regler1.aspx).

Enligt SFIF:s hemsida är den gällande regelboken inom friidrott väldigt omfattande. Ett nytt rekord erkänns officiellt endast då det har presterats enligt dessa fastställda regler och tillåtna förhållanden. I största möjliga mån ska samma förutsättningar och villkor gälla för alla mästerskap världen över, varför det är absolut nödvändigt att det finns tydliga definitioner och specifikationer kring bedömningen av prestationerna. För att löparnas prestationer vid olika tillfällen och tävlingar ska kunna jämföras på ett rättvist och meningsfullt sätt eftersträvas så pass standardiserade resultat som möjligt. Det är självfallet absolut nödvändigt att resultaten ska kunna vägas mot varandra för att det ska finnas någon mening med att sätta rekord.

1.3 Syfte och metod

Syftet med detta arbete är att undersöka hur olika faktorer inverkar på rekordtider i löpning. Vilka faktorer har den största inverkan på löparnas prestationer och hur kan denna beskrivas?

För att undersöka de olika faktorernas effekter tas en lämplig statistisk modellering fram som på bästa möjliga sätt beskriver datamaterialet. Fokus kommer att ligga på olika utföranden av linjära regressionsmodeller för att identifiera och gradera olika verksamma variationskällor. De diagnostiska metoder som kommer att tillämpas för modellkritik är förklaringsgrad, antal variabler, variansskattning, residualanalys och signifikansen för de individuella variablerna. Programpaketet SAS kommer att användas för samtliga beräkningar och grafer i arbetet.

2 Beskrivning av datamaterialet

2.1 Insamling av data

Det datamaterial som arbetet baseras på 124 aktuella rekord i löpning; dels 62 världsrekord hämtade från IAAF:s hemsida (se www.iaaf.org) och dels 62 svenska rekord hämtade från SFIF:s hemsida (se www.friidrott.se). På dessa hemsidor publiceras listor löpande över de senast gällande officiellt godkända rekorden. En begränsning görs till att undersöka rekord på distanserna 50-10 000 meter. Det datamaterial som analysen baseras på hämtades slutgiltigt 2011-08-30 (se Tabell A1-A4 i Appendix).

2.2 Beskrivning av variabler

Den statistiska analysen av datamaterialet kommer att göras på fem utvalda faktorer som på olika sätt kan förväntas förklara en löpares prestation: distans, kön, arenatyp, åldersgrupp och rekordtyp. Responsvariabeln är rekordtiden i sekunder. Syftet med undersökningarna är följaktligen att fastställa hur uppsättningen av olika faktorer inverkar på responsvariabeln tid.

Informationen för samtliga variabler samlas in enligt den rekordstatistik som finns publicerad på IAAF:s och SFIF:s hemsidor. En översiktlig beskrivning av variablerna följer nedan.

Tabell 1

Variabel	Antal	Beskrivning
tid	124	5.56 – 2105.30 sekunder
distans	13	50 - 10 000 meter
kön	2	man eller kvinna
arenatyp	2	inomhus eller utomhus
åldersgrupp	2	junior eller senior
rekordtyp	2	världsrekord eller svenskt rekord

Beskrivning av variabler.

Det bör påpekas att värdena på variabeln distans är ojämnt fördelade mellan de två nivåerna hos kategorivariablerna åldersgrupp respektive arenatyp. Det finns t.ex. endast observationer för juniorer som löper på utomhusarenor. Vidare finns den längsta distansen 10 000 meter endast representerad för utomhusarenor medan de två kortaste distanserna 50 och 60 meter å andra sidan endast har observationer för inomhusarenor. Den ojämna fördelningen bör tas i särskilt beaktande i den kommande analysen för att undvika felaktiga resultat och slutsatser.

3 Metoder och begrepp i arbetet

3.1 Regressionsanalys

Regressionsanalys är en statistisk metod som används för att studera samband mellan en responsvariabel och en eller flera förklaringsvariabler samt en slumpterm. Målet är att anpassa en modell efter vad som bäst passar den aktuella uppsättningen observationer, vilket vanligen görs med minstakvadratmetoden. Genom att variera en variabel i taget i regressionsmodellen medan övriga hålls konstanta kan variablernas respektive effekt på responsvariabeln studeras.

Vid metoden linjär regression utgår man från att en linjär modell kan beskriva sambandet mellan de aktuella variablerna. Följande avsnitt ger en mer utförlig beskrivning av de två verktygen enkel linjär regressionsmodell och multipel linjär regressionsmodell med parametrar enligt definition i Rolf Sundbergs kompendium Lineära Statistiska Modeller.

3.1.1 Enkel linjär regressionsmodell

Modellen vid enkel linjär regression definieras enligt

$$Y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

där Y_i är responsvariabel, x_i är förklaringsvariabel, α och β är parametrar, ε_i är en slumpterm, och $i = 1, \dots, N$ där N är antalet observationer. Modellen förutsätter att slumptermerna är sinsemellan oberoende och normalfördelade enligt $N(0, \sigma^2)$.

Regressionsmodellen har en väntevärdesstruktur enligt

$$\mu(x_i) = \alpha + \beta \cdot x_i$$

där samma definitioner gäller som ovan.

3.1.2 Multipel linjär regressionsmodell

Den enkla linjära regressionsmodellen kan utvidgas till en multipel linjär regressionsmodell i de fall då flera olika förklaringsvariabler kan tänkas ha en effekt på responsvariabeln.

Modellen vid multipel linjär regression definieras enligt

$$Y_i = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \cdots + \beta_n \cdot x_{ni} + \varepsilon_i$$

där Y_i är responsvariabel, $x_{1i}, x_{2i}, \dots, x_{pi}$ är förklaringsvariabler, $\alpha, \beta_1, \dots, \beta_n$ är parametrar, ε_i är en slumpterm, och $i = 1, \dots, N$ där N är antalet observationer. Modellen förutsätter att slumptermerna är sinsemellan oberoende och normalfördelade enligt $N(0, \sigma^2)$.

Regressionsmodellen har en väntevärdesstruktur enligt

$$\mu(x_i) = \alpha + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n$$

där samma definitioner gäller som ovan.

Det kan särskilt nämnas att förklaringsvariablerna som analyseras med hjälp av en multipel linjär regressionsanalys kan vara 0,1-dummyvariabler. Sådana dummyvariabler kan användas för att markera och på så sätt enkelt kunna göra jämförelser av olika kvalitativa kategorinivåer i datamaterialet, se avsnitt 3.3.

3.2 Heteroskedasticitet

Med begreppet heteroskedasticitet menas att storleken på residualernas varians inte är konstant utan istället varierar proportionellt med antingen kovariaten eller någon kategorivariabel. Det kan medföra att olika statistiska tekniker som används för att till exempel titta på hypotestest och standardavvikelser blir inkorrekta och otillförlitliga.

Som det nämndes i avsnittet ovan om linjära regressionsmodeller är det en förutsättning för regressionsanalysen att slumptermerna kan antas oberoende och normalfördelade med en konstant varians, varför det är ett problem om heteroskedasticitet förekommer i modellen.

3.3 Dummyvariabel

En dummyvariabel är en variabel som används vid regressionsanalys för att markera olika kvalitativa kategorinivåer som observationerna är uppdelade i. Det är möjligt att ha flera olika koder för en dummyvariabel, men i de allra enklaste fallen representeras nivåerna av en så kallad 0,1-dummyvariabel som endast kan anta ett av de två värdena 0 eller 1. Genom att använda dummyvariabler i regressionsmodeller kan man på ett överskådligt sätt representera olika delmängder av datamaterialet utan att behöva skapa separata modeller för varje enskild egenskap. Till exempel kan en dummyvariabel användas för att låta representera effekten av olika kön genom att låta den anta värdet 0 för män och 1 för kvinnor. Denna representerar då skillnaden i intercept mellan de olika könen. Genom att utvidga regressionsmodellen till att även inkludera en produktterm mellan dummyvariabeln och den aktuella förklaringsvariabeln kan det även testas huruvida det de olika könen har olika lutning, vilket skulle innebära att faktorn kön modifierar effekten av förklaringsvariabeln. På detta sätt kan man på ett tämligen okomplicerat sätt inkludera och undersöka effekter av olika kategorier på responsvariabeln.

3.4 Förklaringsgrad och justerad förklaringsgrad

Förklaringsgraden R^2 definieras som den andel av den totala variationen i responsvariabeln som förklaras av en bestämd linjär modell och är ett anpassningsmått som ofta används vid regressionsanalys för att jämföra olika modeller med varandra.

Andelen förklarad variation ges av kvoten

$$R^2 = \frac{Kvs(regression)}{Kvs(totalt)} = 1 - \frac{Kvs(residual)}{Kvs(totalt)}$$

Förklaringsgraden är ett tal mellan 0 och 1 som anger hur pass väl en specifik regressionslinje approximerar de aktuella observationerna och därigenom hur pass väl modellen lämpar sig för att beskriva sambandet mellan variablerna. Ju högre förklaringsgrad, desto starkare är det linjära sambandet och mer variation i data lyckas således förklaras av den skattade modellen.

En nackdel med förklaringsgraden är att dess värde alltid ökar när ytterligare variabler läggs till i en modell och det även om den införda variabeln skulle vara rent slumpmässig. Det beror på att *Kvs(residual)* då inte kan minska i storlek. Ju fler variabler som läggs till i en modell, desto större blir per automatik dess förklaringsgrad och det utan att det behöver innebära att variablerna har någon som helst statistisk signifikans. Av den anledningen kan det vid jämförelse av multipla linjära regressionsmodeller vara ett bättre alternativ att använda den justerade förklaringsgraden R^2_{adj}, vilken justerar för antalet förklaringsvariabler i modellen. Den justerade förklaringsgraden mäter hur mycket variansreduktion som har uppnåtts genom att föra in en variabel i en modell baserat på huruvida den ger en mindre variansskattning $\widehat{\sigma^2}$, vilket inträffar då variabeln förbättrar modellen mer än vad som kan förväntas av slumpen.

Den justerade förklaringsgraden fås enligt

$$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = 1 - \frac{Kvs(residual)/(N-k-1)}{Kvs(totalt)/(N-1)}$$

Här är *N* antalet observationer i analysen och *k* antalet förklaringsvariabler i modellen. Vi kan betrakta $\hat{\sigma}_0^2$ som den σ^2-skattning som fås när ingen förklaringsvariabel ingår i modellen. Den justerade förklaringsgraden R^2_{adj} fås lika med eller mindre än R^2.

3.5 Stegvis variabelselektion

Vid en multipel linjär regressionsanalys kan det vara en svår uppgift att välja ut den delmängd ur en större uppsättning förklaringsvariabler som har den största effekten på responsvariabeln. Genom att använda stegvis variabelselektion väljs de variabler som tillsammans utgör den bästa regressionsmodellen ut baserat på deras signifikansnivåer. Vid varje steg i proceduren antingen inkluderas eller elimineras successivt en variabel för att få en så bra anpassning som möjligt fram till det att ett visst stoppkriterium uppfylls. Metoder för automatisk stegvis variabelselektion besparar oss en hel del beräkningsarbete, men det finns för den delen inget som garanterar att slutresultatet är det mest optimala för observationerna. Man bör snarare se det som att proceduren tillhandahåller signifikanta modeller, vilka man sedan kan använda som utgångspunkt för fortsatta undersökningar och analyser av de ingående variablerna.

I detta arbete kommer vi att använda oss av följande tre stegvisa procedurer:

Backward elimination
Proceduren utgår från en modell innefattande samtliga variabler och utesluter sedan successivt en variabel i taget fram till det att proceduren stannar. Hypotesen $\beta_h = 0$ testas i varje steg för alla kvarvarande variabler x_h på en i förhand vald signifikansnivå. Om en eller flera variabler i modellen inte fås som signifikant skilda från noll elimineras den variabel som ger den minsta sänkningen av förklaringsgraden R^2. Proceduren stannar när alla kvarvarande parametrar β_h i modellen fås som signifikant skilda från noll.

Forward selection
Proceduren utgår från en modell innefattandes endast ett intercept och inga variabler. Sedan utvidgas modellen successivt med en variabel i taget genom test av hypotesen $\beta_h = 0$ för alla kvarvarande variabler x_h. Den variabel som fås mest signifikant i varje steg inkluderas i modellen så länge som det finns minst en variabel kvar som är signifikant på en i förhand vald signifikansnivå.

Stepwise regression
Proceduren kan ses som en kombination av de två procedurerna Backward elimination och Forward selection. Den utgår från en modell innefattandes endast ett intercept och inkluderar sedan successivt en variabel i taget baserat på signifikansnivå. Efter varje steg kontrolleras dessutom variablerna som har inkluderats i modellen för att se om de fortfarande fås som signifikanta, om inte så elimineras dem. Detta tillvägagångssätt motiveras av att variabler kan bli överflödiga när ytterligare variabler som förklarar samma del av variationen läggs till i modellen. Proceduren stannar när ingen mer inkludering eller exkludering sker.

3.6 Log-transformation

I de fall då $\mu(x)$ enligt definitionen i avsnitt 3.1.1 inte är linjär i sina parametrar α och β bör det inför tillämpning av linjär regression som förutsätter linjära samband undersökas huruvida man med hjälp av en lämplig transformation kan överföra modellen till en (approximativt) linjär form. Tillämpning av log-transformation är inte bara en gynnsam metod för att i dessa fall korrigera icke-linjäritet, utan kan även användas för att minska problem med observerad heteroskedasticitet eller få en snedfördelad variabel mer normalfördelad.

En multiplikativ modell kan definieras enligt

$$Y = \mu(1 + \varepsilon), \mu(x) = \alpha \cdot x^{\beta}$$

där $x > 0, \alpha > 0$ och $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. Genom att tillämpa metoden log-transformation på modellens Y-värden fås som resultat en (approximativt) linjär modell enligt

$$Y' = \log(Y) = \alpha' + \beta \cdot \log(x) + \varepsilon'$$

där $\alpha' = log(\alpha)$ och $\varepsilon' = log(1 + \varepsilon) \approx \varepsilon$.

Modellerna ovan åskådliggör hur man med en log-transformation får en multiplikativ modell att uttryckas på linjär form och därmed möjliggör anpassning med hjälp av linjär regression.

3.7 Variansanalys (ANOVA)

En nollhypotes formulerad som en linjär hypotesmodell enligt definitioner i avsnitt 3.1.1 och 3.1.2 kan prövas inom ramen för en större grundmodell som förutsätts vara giltig genom att genomföra en variansanalys, där de olika verksamma variationskällorna i datat identifieras och jämförs. Resultatet som fås av en variansanalys rapporteras ofta i en sammanfattande och överskådlig ANOVA-tabell.

En generell ANOVA-tabell redovisas nedan med beteckningar enligt Rolf Sundbergs definition i kompendiet i Lineära Statistiska Modeller.

Tabell 2

Variationskälla	**Fg**	**Kvs**	**Mkvs**	**Väntevärde för Mkvs**
Avvikelse från hypotesmodellen	$k-l$	$\left\|\hat{\mu}-\hat{\hat{\mu}}\right\|^2$	$\frac{\left\|\hat{\mu}-\hat{\hat{\mu}}\right\|^2}{k-l}$	$\sigma^2+\frac{\left\|\mu-\hat{\hat{\mu}}\right\|^2}{k-l}$
Residual inom grundmodellen	$N-k$	$\|Y-\hat{\mu}\|^2$	$\hat{\sigma}^2$	σ^2
Residual inom hypotesmodellen	$N-l$	$\left\|Y-\hat{\hat{\mu}}\right\|^2$		

Generell ANOVA-tabell.

Här är N antalet observationer i analysen, k antalet parametrar i grundmodellen och l antalet parametrar i hypotesmodellen ($l < k$). I tabellen representerar $\hat{\mu}$ minstakvadratskattningen av μ i grundmodellen och $\hat{\hat{\mu}}$ minstakvadratskattningen i hypotesmodellen. Vidare betecknar $\| \|^2$ skalärprodukten av en vektor med sig själv och längden (normen) för en vektor definieras som kvadratroten ur denna skalärprodukt med beteckningen $\| \|$.

Vid test av en nollhypotes kan man genom variansanalys undersöka hur variationen omkring den aktuella hypotesmodellen delas upp i de två komponenterna residual inom grundmodellen $\|y-\hat{\mu}\|^2$ respektive avvikelse från hypotesmodellen inom grundmodellen $\left\|\hat{\mu}-\hat{\hat{\mu}}\right\|^2$.

Ett test på signifikansnivå p av den linjära nollhypotesen för ett test fås med hjälp av kvoten mellan medelkvadratsummorna genom att hypotesen förkastas då

$$\frac{\frac{\left\|\hat{\mu}-\hat{\hat{\mu}}\right\|^2}{k-l}}{\hat{\sigma}^2} > F_p(k-l, N-k)$$

3.8 Kovariansanalys (ANCOVA)

Kovariansanalys är motsvarigheten till en vanlig variansanalys när förklaringsvariablerna är en blandning av kovariat och kategorivariabler. Det är en metod som används för att kunna studera kategorivariablernas effekter i närvaro av en kovariat som vi vill kontrollera för. Genom att hänsyn tas till de olika kategorivariablerna i data kan man öka den statistiska

styrkan i analysen då dessa redogör för en del av variationen i responsvariabeln och på så sätt minska residualernas variation. Med hjälp av ANCOVA kan man genom att testa en serie av hypoteser jämföra och dra slutsatser kring olika kategorinivåers linjära regressionsmodeller.

För en envägs-ANCOVA då det finns multipla kovariater för var och en av de aktuella observationerna används en grundläggande modell, vilken enligt boken *Analysis of Messy Data, Volume III: Analysis of Covariance* (Milliken & Johnson, 2002) definieras enligt

$$Y_{ij} = \alpha_i + \beta_{1i} \cdot x_{1ij} + \cdots + \beta_{ki} \cdot x_{kij} + \varepsilon_{ij}$$

Här är Y_{ij} responsvariabel, x_{pij} den p:te kovariatens värde för den ij:te observationen, α_i intercept för den i:te kategorigruppens regressionsyta, β_{pi} lutning för den p:te kovariatens riktning för kategorigrupp i, ε_{ij} en slumpterm, och j=1,…n, i=1,…,t. För alla slumptermer förutsätter modellen att de är sinsemellan oberoende och normalfördelade enligt $\mathrm{N}(0, \sigma^2)$.

Var och en av de olika kategorinivåerna beskrivs enligt en multipel linjär regressionsmodell och kan alltså ha många olika utseenden. Det kan vara intressant att inkludera samspelstermer mellan de olika förklaringsvariablerna i modellen, då utgångspunkten att de olika linjerna är parallella enligt ovan inte alltid är realistisk. Genom att multiplicera två eller fler variabler fås en samspelsterm som kan användas för att analysera effekterna av de respektive variabler som ingår. En utvidgad modell av detta slag tillåter att både intercept och lutningar skiljer sig åt mellan kategorinivåerna, vilket därmed möjliggör nivåspecifika linjära regressionsmodeller.

Det första steget i en ANCOVA är att beräkna en regressionslinje för var och en av de olika kategorinivåerna (eller kombinationer av kategorinivåer). Sedan testas huruvida de olika linjernas lutningar skiljer sig signifikant åt. Om så är fallet innebär det att linjerna korsar varandra i någon punkt på grafen och att man inte kommer längre med hjälp av ANCOVA. Slutsatsen blir då att det finns en skillnad i lutning mellan de olika kategorinivåerna, vilket innebär att kovariaten har olika effekt på responsvariabeln beroende på observationernas kategoritillhörighet. Om hypotesen att linjerna har samma lutning å andra sidan inte fås som signifikant kan det antas att linjerna är parallella. Man övergår då till att testa huruvida linjerna skiljer sig åt med avseende på sina respektive intercept. Eftersom att linjerna antas parallella är skillnaden i y-led mellan dem konstant för alla värden på kovariaten, vilken därmed enkelt kan bestämmas genom att jämföra intercepten. Finns det ingen signifikant skillnad mellan de olika kategorinivåernas intercept kan det antas att de beskrivs enligt en och samma modell, det vill säga att det inte finns någon statistiskt signifikant skillnad mellan nivåerna. Om det anses relevant i den aktuella studien kan man med hjälp av ANCOVA även testa hypoteserna att minst en av regressionslinjernas intercept respektive lutning är noll.

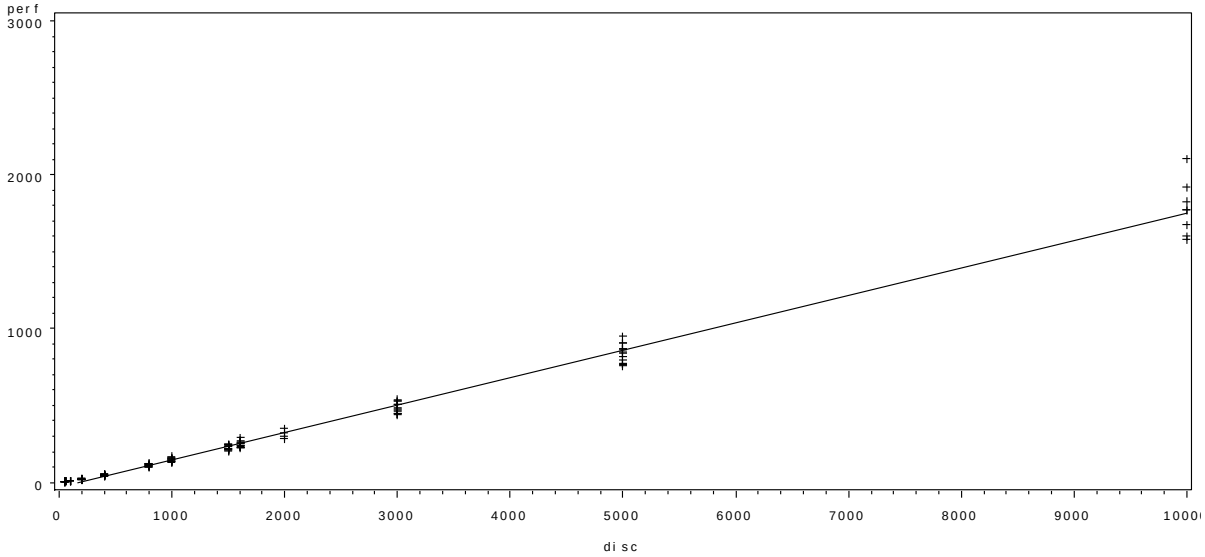
4 Statistisk modellering och variansanalys

4.1 Undersökning av datamaterialet

Det ursprungliga datamaterialet består av statistik över totalt 124 rekord. En initial och viktig del av bedömningen av variablernas effekt på de observerade rekorden kan göras illustrativt genom att studera olika scatterplots. Det ger oss en översikt av sambanden mellan de olika variablerna som ska undersökas och gör att eventuellt avvikande observationer tidigt kan upptäckas. Dock är det viktigt att det som tidigare nämnts i avsnitt 2.2 rörande datats ojämna fördelning mellan kategorinivåerna has i åtanke för att inga felaktiga slutsatser ska dras.

En scatterplot över sambandet mellan variablerna tid och distans presenteras enligt nedan. En regressionslinje som har anpassats till observationerna enligt minstakvadratmetoden inkluderas i grafen för att sambandet mellan variablerna enklare ska kunna studeras.

Figur 1



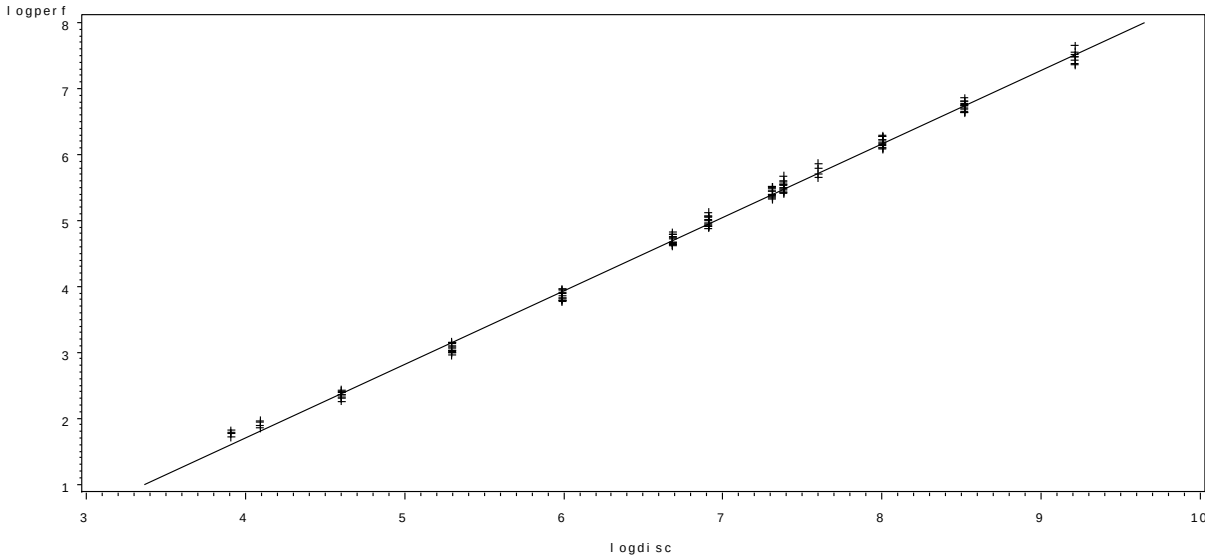
Scatterplot: tid i förhållande till distans.

Som väntat finns det ett starkt linjärt beroende mellan variablerna distans och tid. Korrelationen mellan dem är starkt positiv med en mycket hög korrelationskoefficient som $r = 0.9940$. Alltså förklarar distans på egen hand en absolut majoritet av den variation som finns i tid, vilket stämmer väl överrens med vad som förutspåddes i inledningen. Vidare ser man tydligt att observationernas spridning kring den räta regressionslinjen ökar proportionellt med växande värden på distans. Det tyder på att residualernas varians inte fås som konstant, utan att det förekommer heteroskedasticitet. Det kan även observeras att sambandet mellan variablerna inte ser ut att beskrivas på linjär form då en dålig anpassning fås mellan linje och observationer. För de kortaste och längsta distanserna fås en underskattning av tiderna medan vi för distansen på 5000 meter i grafens mitt tvärtemot får en överskattning. Det kan förklaras av att det hos observationerna finns systematiska avvikelser från linjäritet i form av en svagt

konvex krökning. Att rekordtiderna istället fås utmed en potenskurva antas bero på att löpare som springer de kortare distanserna lyckas hålla en högre hastighet genom hela loppet medan löpare som springer de längre distanserna inte klarar av att hålla sin maximala hastighet från start till mål. I övrigt går det inte att i scatterploten identifiera några avvikande observationer.

Baserat på de gjorda iakttagelserna anses variablerna tid och distans vara i behov av lämplig transformation inför kommande tillämpningar av linjär regression, både för att förbättra antagandet om linjäritet och för att försöka råda bot på den observerade heteroskedasticiteten. En lämplig transformation för just dessa ändamål är den så kallade log-transformationen (se avsnitt 3.6). Vi log-transformerar variablerna och undersöker huruvida det ger önskad effekt genom att studera en scatterplot över sambandet mellan de alternativa variablerna log(tid) och log(distans). Vi inkluderar även här en regressionslinje som har anpassats till observationerna enligt minstakvadratmetoden.

Figur 2

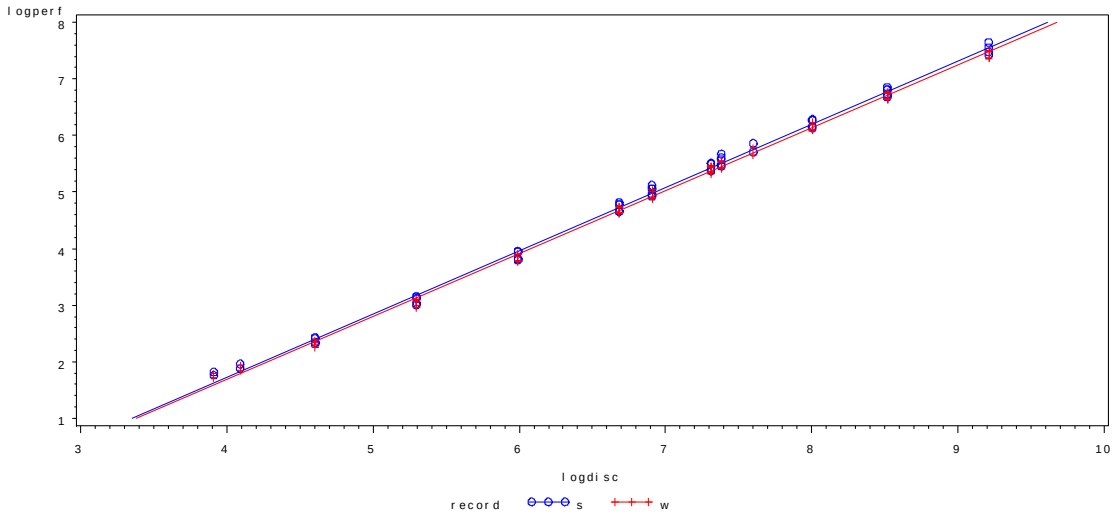


Scatterplot: log(tid) i förhållande till log(distans).

I scatterploten ser vi att log-transformationen verkar ha lyckats åstadkomma ett betydligt mer linjärt samband mellan variablerna, samtidigt som problemet med heteroskedasticitet upphör. Därmed kan slutsatsen dras att de log-transformerade observationerna approximativt uppfyller de modellantaganden som behöver vara uppfyllda vid en linjär regression och således är dessa att föredra i den kommande regressionsanalysen. Fortsättningsvis kommer vi därför att istället utgå från det log-transformerade datamaterialet i våra undersökningar.

Det datamaterial som studeras består av fyra kategorivariabler med två olika nivåer. För att få en bild av hur det log-transformerade datat varierar inom respektive kategori studerar vi deras scatterplots med en utritad linjär regressionslinje för respektive kategorinivå.

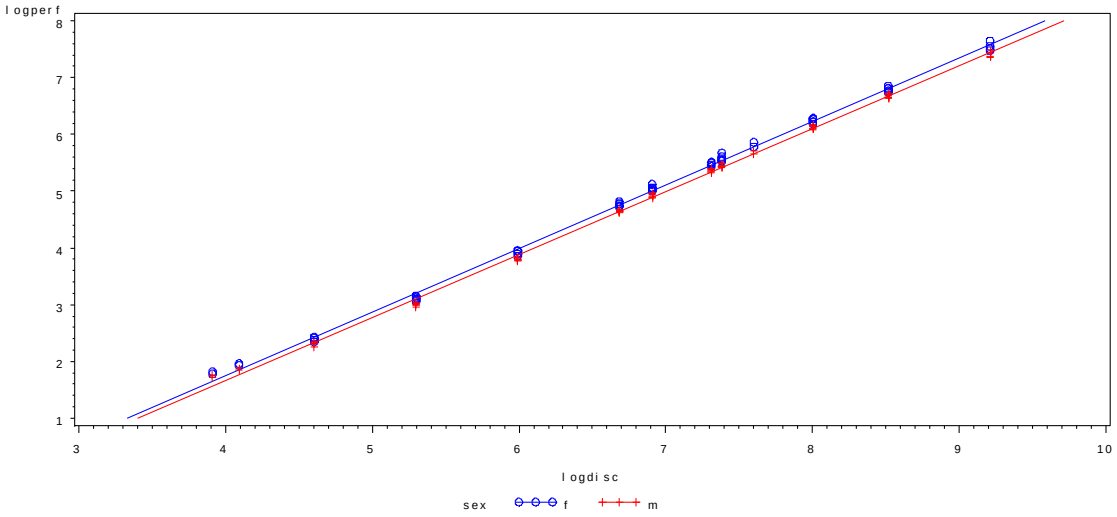
Figur 3



Scatterplot: log(tid) i förhållande till log(distans) för kategori rekordtyp.

Regressionslinjen för svenska rekord ligger över den för världsrekord för samtliga värden på log(distans), vilket säger oss att de svenska rekordtiderna i datat är genomgående sämre. Den observerade skillnaden ser dessutom ut att öka med växande värden på distans, vilket innebär att skillnaden i prestation mellan rekordtyperna blir särskilt märkbar vid långdistanslöpning.

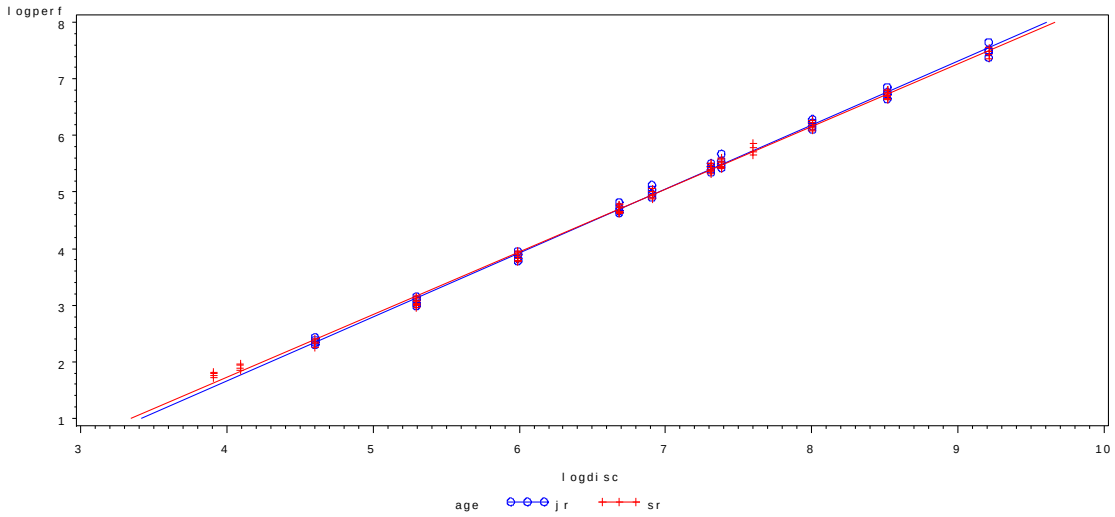
Figur 4



Scatterplot: log(tid) i förhållande till log(distans) för kategori kön.

Män får genomgående bättre rekordtider än kvinnor. Av samtliga scatterplots som görs för de olika kategorierna observeras här den största skillnaden i y-led mellan regressionslinjerna för respektive nivå. De två regressionslinjerna ser ut att fås ungefärligt parallella, vilket tyder på att könen har samma utveckling av sina tider.

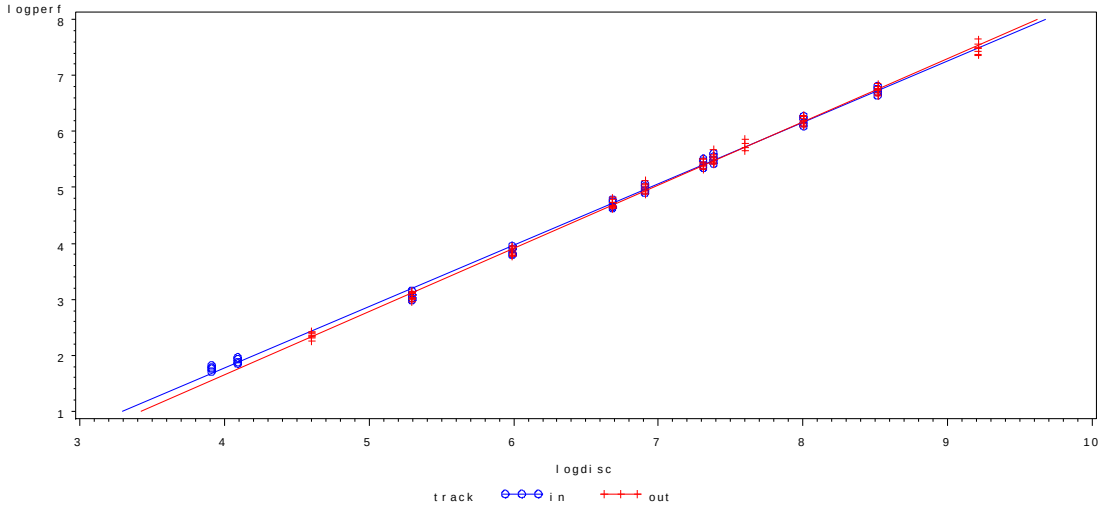
Figur 5



Scatterplot: log(tid) i förhållande till log(distans) för kategori åldersgrupp.

Den ojämna fördelningen av observationer fås särskilt stor mellan de olika nivåerna inom kategorin åldersgrupp. Nivån junior representerar endast 40 av totalt 124 observationerna i datamaterialet och dessutom finns endast rekordtider för lopp på utomhusarenor. En sådan skevhet i data kan lätt leda till att feltolkningar görs när man studerar endast en variabel i taget på det här sättet, varför inte allt för stor vikt bör läggas vid utseendet av denna scatterplot.

Figur 6



Scatterplot: log(tid) i förhållande till log(distans) för kategori arenatyp.

Även för arenatyp fås en ojämn fördelning mellan de olika nivåerna och det är även här svårt att bilda sig en uppfattning om eventuella skillnader. Det noteras att regressionslinjerna skär varandra i scatterplotens mitt och att lopp utomhus står för de bästa prestationerna för de kortare distanserna. De två kortaste distanserna har endast rekord för löpare på inomhusarenor

och den längsta distansen har endast rekord för utomhusarenor, vilket gör det mycket troligt att vi får dåligt anpassade regressionslinjer för dessa nivåer.

Inga uppenbara outliers kan ses i någon av scatterplotsen ovan. Vi avstår i detta avsnitt från att göra jämförelser av observationernas spridningar för fixerade värden på log(distans) då den till stor del inte kan anses vara slumpmässig utan istället beror på de övriga faktorerna.

Inför tillämpning av linjär regression är främst upptäckterna av systematiska avvikelser från linjäritet och heteroskedasticitet i datamaterialet av stor betydelse. Avvikelserna motiverar att den linjära regressionsanalysen istället bör genomföras på de log-transformerade variablerna. Vi har genom att studera scatterplots för de olika kategorierna kunnat observera en påtaglig skillnad mellan nivåerna inom kön respektive rekordtyp. Däremot var det svårare att dra några slutsatser ur scatterplotsen för kategorierna åldersgrupp och arenatyp på grund av den påtagligt skeva fördelningen av rekord.

4.2 Regressions- och variansanalys

Innan tillämpningen av regressionsanalys på datamaterialet finns det en effekt som bör ses över för att de resulterande modellerna ska fås så pass korrekta och lättolkade som möjligt, nämligen effekten av reaktionstid. Med reaktionstid menas här den tid det tar från det att startsignalen når löparens öra tills det att denne reagerar och loppet startar. Enligt regelverket räknas en start i löpning som en tjuvstart då loppet påbörjas tidigare än 0.1 sekunder efter startskottet (se http://www.friidrott.se/alltom/regler/regler2.aspx). Därmed kan vi betrakta denna tid som en ren reaktionstid vars effekt enkelt kan kompenseras för genom att subtrahera 0.1 sekunder från samtliga tider i rekordstatistiken. Efter detta medelvärdescentreras tidsvariabeln för att underlätta inför kommande tolkningar av linjära regressionsmodeller.

Regressionsanalysen inleds med en enkel linjär regressionsmodell med den ursprungliga förklaringsvariabeln distans för att se hur pass stor del av variationen i responsvariabeln tid som den ensamt kan förklara. En variansanalys av modellen ger följande ANOVA.

Tabell 3

Variationskälla	Fg	Kvs	Mkvs	F	R^2/R^2_{adj}
Regression	1	25195440	25195440	10001.9	0.9879/0.9879
Residual	122	307326	2519.0644		
Total	123	25502766			

ANOVA: förklaringsvariabel distans.

Den höga justerade förklaringsgraden visar att distans ensamt förklarar en betydande del av variationen i tid, närmare bestämt 98.79 %. Av detta drar vi slutsatsen att distans absolut bör ingå i den modell som önskas beskriva observationerna så bra som möjligt.

De tidigare undersökningarna av datat i avsnitt 4.1 visade att sambandet mellan tid och distans inte fås på linjär form, utan istället kan beskrivas enligt en svagt konvex potensfunktion.

Vi utgår följaktligen från att det ursprungliga datamaterialet kan beskrivas enligt en funktion på formen

$$tid_i = \alpha \cdot distans_i^{\beta} \cdot \varepsilon_i$$

Genom att använda log-transformation på dessa variabler överförs sambandet istället på linjär form enligt funktionen

$$log(tid_i) = log(\alpha) + \beta \cdot log(distans_i) + \varepsilon_i$$

Det linjära sambandet möjliggör en korrekt modellering med metoden linjär regression och vid en ny tillämpning av enkel linjär regressionsanalys för de alternativa log-transformerade variablerna fås följande ANOVA.

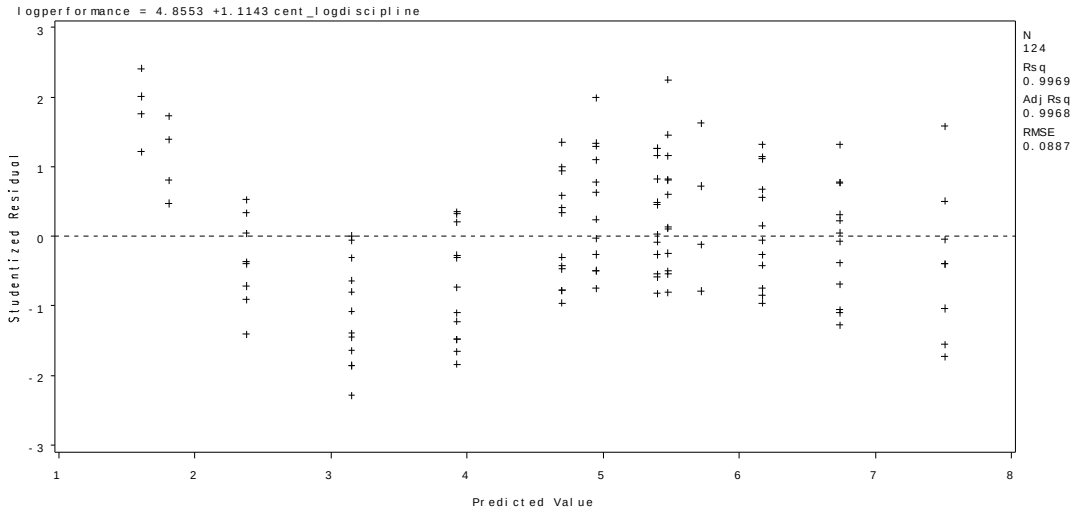
Tabell 4

Variationskälla	**Fg**	**Kvs**	**Mkvs**	**F**	$\mathbf{R^2/R^2_{adj}}$
Regression	1	304.7887	304.7887	38745.1	0.9969/0.9968
Residual	122	0.9597	0.0079		
Total	123	305.7484			

ANOVA: förklaringsvariabel log(distans).

Den redan utmärkt höga justerade förklaringsgraden stiger från 98.79 % upp till hela 99.68 %, vilket med tanke på den smärre andelen återstående oförklarad variation är en betydande höjning. Modellens parameterskattning $\hat{\beta} = 1.1143$ stämmer väl överrens med antagandet om en ursprunglig svagt konvex potensfunktion. För att vidare studera sambandet mellan log(tid) och log(distans) studeras en residualplot över modellens standardiserade residualer.

Figur 7



Residualplot: standardiserade residualer i förhållande till predikterade värden.

Att döma av residualernas spridning kring nollstrecket har log-transformationen av observationerna resulterat i att det tidigare problemet med heteroskedasticitet nu har reducerats något då variationen har jämnats ut. Däremot visar residualernas fördelning upp en påtaglig systematik, vilket innebär att modellantagandet om oberoende residualer inte är uppfyllt.

De observationer som har de största avvikande residualerna från nollstrecket hör till de två distanserna 50 och 60 meter. Regressionslinjen får alltså en särskilt dålig anpassning för de två kortaste loppen och att de är positiva innebär att de faktiska värdena är större än vad modellen motiverar. En effekt som orsakar att just de kortaste distanserna underskattas av modellen är accelerationstiden, det vill säga den tidsförlust som orsakas av att löparen startar från stillastående läge. För distanserna 50 och 60 meter tar det en större del av loppet för att accelerera till den maximala hastighet som sedan kan till målgång, vilket gör det svårt att utröna själva löpeffekten för dessa observationer. Vidare är den potensfunktion som beskriver sambandet mellan variablerna distans och tid lämplig för att modellera den trötthetseffekt hos löparna, vilken kan observeras i Figur 2 där växande värden på distans ger en avtagande medelhastighet. Denna trötthetseffekt saknas däremot helt för rekorden på de kortaste distanserna, där löparna tvärtom har en lägre hastighet i början än i huvuddelen av loppet. Sammantaget anses alltså effekten av accelerationstiden och avsaknandet av tröttshetseffekt för distanserna 50 och 60 meter till stor del kunna förklara avvikelserna i residualploten ovan.

Det kan ifrågasättas huruvida observationerna för distanserna 50 och 60 meter bör ingå i det utvalda datamaterialet eftersom att det anses vara osäkert om de motsvarar den löpeffekt som vi önskar att studera. Genom att anti-logaritmera de predikterade värdena fås för distansen 50 meter en tidsskattning på $e^{1.6103} \approx 5.00$ sekunder och för distansen 60 meter en tidsskattning på $e^{1.8135} \approx 6.13$ sekunder. Jämförs dessa skattningar med de observerade tiderna ser vi att de predikterade värdena ger en obestridlig underskattning för båda distanser. Vi kan beräkna den faktiska avvikelsen genom att anti-logaritmera residualerna, vilket resulterar i beräknade procentpåslag för distansen 50 meter på 16.50 %, 11.10 %, 23.29 %, 19.10 % och för distansen 60 meter på 7.32 %, 4.22 %, 16.28 %, 12.85 %. Det visar på att reaktionstiden har en betydande procentuell inverkan på tiderna. Det anses motiverat att de åtta inflytelserika observationer tillhörande distanserna 50 och 60 meter bör uteslutas ur datat inför den fortsatta analysen. Ännu ett argument som talar för detta är att observationerna endast finns representerade för den ena nivån inom åldersgrupp (senior) respektive arenatyp (inomhus). Sådana skeva fördelningar ger vid en modellanpassning många gånger missvisande och svårtolkade resultat.

De åtta observationerna tillhörande distanserna 50 och 60 meter utesluts därmed från det log-transformerade datamaterialet, varpå en enkel linjär regressionsanalys tillämpas på samma sätt som tidigare. Resultatet för denna fås enligt en ANOVA nedan.

Tabell 5

Variationskälla	**Fg**	**Kvs**	**Mkvs**	**F**	$\mathbf{R^2/R^2_{adj}}$
Regression	1	227.8958	227.8958	34711.4	0.9967/0.9967
Residual	114	0.7485	0.0066		
Total	115	228.6442			

ANOVA: förklaringsvariabel log(distans).

Den totala variationen i log(tid) har minskat avsevärt till följd av att de åtta observationerna har uteslutits ur datat. Den justerade förklaringsgraden 99.67 % visar att den andel av den totala variationen som förklaras av modellen samtidigt har minskat obetydligt.

Vi fortsätter regressionsanalysen med att undersöka vilka ytterligare variabler som har effekt på log(tid) och till följd därav kan användas för att förklara den återstående oförklarade variationen på 0.33 %. Den enkla linjära regressionsmodellen utvidgas till att även inkludera information om de fyra kategorivariablerna rekordtyp, kön, arenatyp och åldersgrupp. Detta görs genom att låta dem representeras av dummyvariabler som kodas enligt tabellen nedan.

Tabell 6

Kategori	**Värde på ursprunglig variabel**	**Värde på dummyvariabel**
rekordtyp	världsrekord	0
	svenskt rekord	1
kön	man	0
	kvinna	1
åldersgrupp	senior	0
	junior	1
arenatyp	utomhus	0
	inomhus	1

Beskrivning av dummyvariabler.

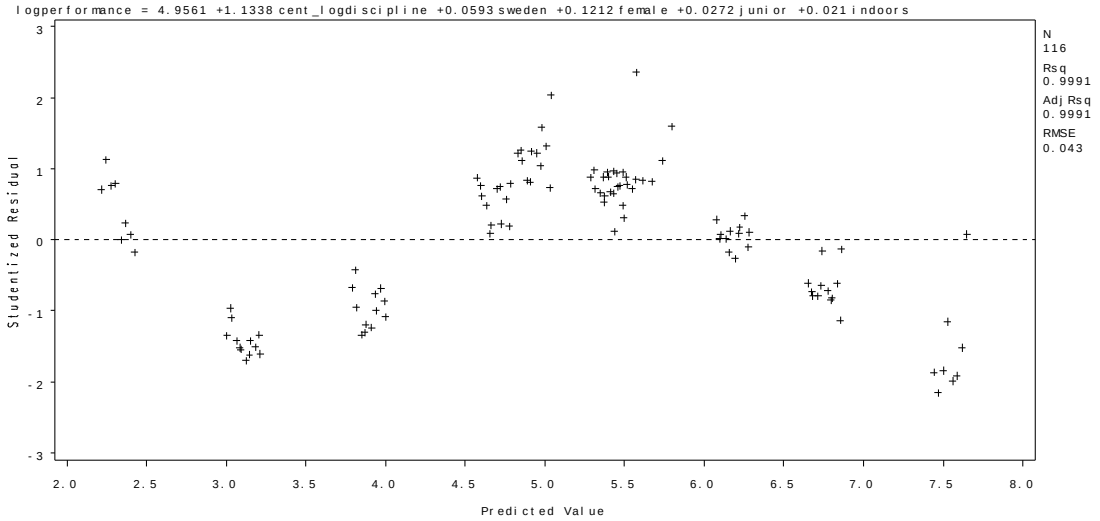
Resultatet av utvidgningen blir en multipel linjär regressionsmodell som innehåller fem förklaringsvariabler: log(distans), rekordtyp, kön, åldersgrupp och arenatyp. En ANOVA för denna modell med tillhörande residualplots över standardiserade residualer fås enligt nedan.

Tabell 7

Variationskälla	**Fg**	**Kvs**	**Mkvs**	**F**	$\mathbf{R^2/R^2_{adj}}$
Regression	5	228.4406	45.6881	24681.6	0.9991/0.9991
Residual	110	0.2036	0.0019		
Total	115	228.6442			

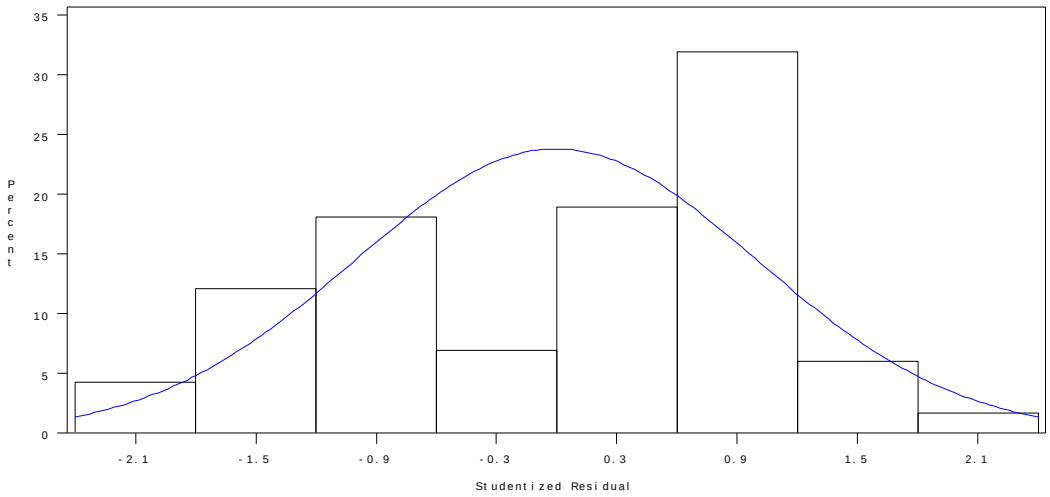
ANOVA: förklaringsvariabler log(distans), rekordtyp, kön, åldersgrupp, arenatyp.

Figur 8



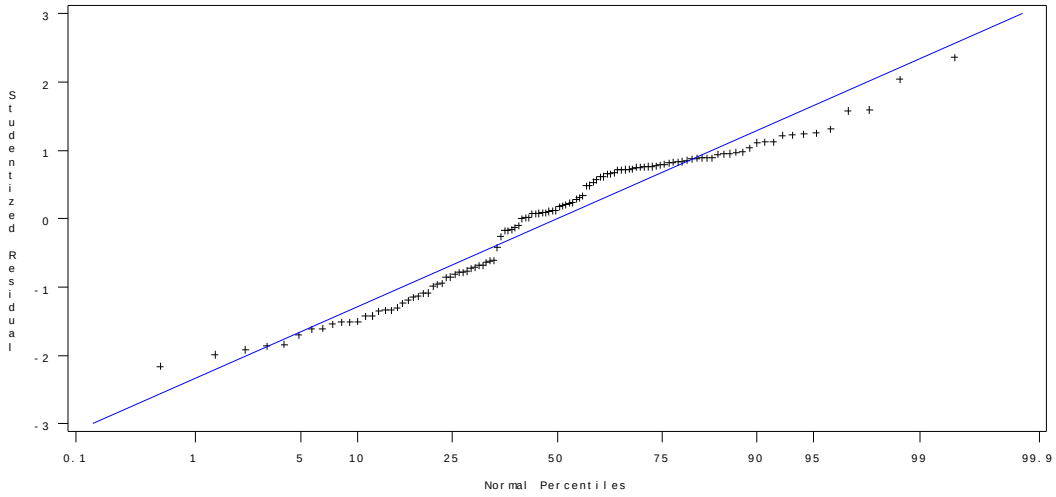
Residualplot: standardiserade residualer i förhållande till predikterade värden.

Figur 9



Residualplot: histogram för standardiserade residualer med normalfördelningskurva.

Figur 10



Residualplot: standardiserade residualer i förhållande till normalpercentilen.

Regressionsmodellen får efter det att dummyvariablerna har inkluderats en högre justerad förklaringsgrad på 99.91 % och även de tillhörande residualplotsen är mer tillfredsställande med avseende på fördelning och slumpmässighet. Det går dock inte att förbise att en viss systematik fortfarande finns kvar i residualerna, vilket tyder på att det finns en underliggande struktur i datat som inte finns representerad i den aktuella modellen. Samtliga variabler i modellen fås statistiskt signifikanta (p-värden ≤ 0.0380). Av dummyvariablerna får kön den överlägset största signifikansen och har därmed största effekten på log(tid).

Signifikant effekter har kunnat påvisas av log(distans) och samtliga kategorivariabler, men för att närmare undersöka hur dessa effekter kan beskrivas fortsätter vi att utvidga vår modell. Systematiken i residualerna skulle kunna tyda på att det finns en viss avvikelse från linjäritet hos observationerna, varför det undersöks huruvida det går att fånga upp en sådan effekt i modellen genom att inkludera potenser av log(distans). Regressionsmodellen utvidgas med de tre förklaringsvariablerna $\log(\text{distans})^2$, $\log(\text{distans})^3$ och $\log(\text{distans})^4$. En variansanalys för modellen ger att samtliga förklaringsvariabler är klart signifikanta (p-värden <.0001) och en förträffligt hög justerad förklaringsgrad fås på 99.99 %. Residualplotsen visar att vi har lyckats råda bot på en stor del av den tidigare observerade systematiken. Slutsatsen dras därför att det med en fjärdegradsmodell är möjligt att i viss mån modellera det svagt icke-linjära sambandet mellan log(distans) och log(tid).

Ytterliggare två utvidgningar görs av regressionsmodellen genom att i första steget inkludera effekterna av dummyvariablernas respektive produkter med log(distans) och i andra steget även inkludera effekterna av dummyvariablernas respektive produkter med varandra. Dessa samspelstermer representerar de interaktioner som bedöms vara relevant att testa i modellen. Båda modellerna får samma höga justerade förklaringsgrad som tidigare på 99.99 % och de tillhörande residualerna uppvisar ett tillfredsställande slumpmässigt beteende. Då en variansanalys genomförs för den första utvidgningen fås samtliga variabler som statistiskt signifikanta (p-värden ≤ 0.0120) med undantag för två samspelstermer: log(distans)·arenatyp (p-värde 0.3098) och log(distans)·åldersgrupp (p-värde 0.3545). Då en variansanalys genomförs för den andra utvidgningen fås en majoritet av variablerna som signifikanta

(p-värden ≤0.0076), men det går inte att påvisa någon effekt av sex samspelstermer: kön·arenatyp (p-värde 0.6920), kön·åldersgrupp (p-värde 0.5784), rekordtyp·arenatyp (p-värde 0.0900), rekordtyp·åldersgrupp (p-värde 0.1102), log(distans)·arenatyp (p-värde 0.2809) och log(distans)·åldersgrupp (p-värde 0.3255). Det kan särskilt observeras att de två interaktionerna log(distans)·arenatyp och log(distans)·åldersgrupp saknar statistiskt påvisbar signifikans i båda de undersökta regressionsmodellerna, men då de trots detta i båda fallen får relativt låga p-värden kan det ifrågasättas huruvida effekterna kan anses vara helt försumbara.

De två successiva utvidgningarna av modellen visar att det eventuellt finns vissa signifikanta samspelstermer att ta i beaktande, men då effekten av flertalet av variablerna inte kan visas bero på annat än slumpmässig variation bör modellen absolut reduceras. Med utgångspunkt från de sammanlagt 18 variabler som nu ingår i modellen vill vi försöka skapa en så pass enkel modell som möjligt utan att för den delen förlora för mycket justerad förklaringsgrad. Problemet med att välja ut en uppsättning av relevanta variabler för detta ändamål kan underlättas genom användning av de tre stegvisa variabelselektionerna Backward elimination, Forward selection och Stepwise regression. Tillämpningen av automatisk variabelselektion för att välja ut signifikanta variabler görs här på en 1 % -ig signifikansnivå. Den uppsättning variabler som metoden utgår från är följaktligen log(distans), $\log(\text{distans})^2$, $\log(\text{distans})^3$, $\log(\text{distans})^4$, rekordtyp, kön, åldersgrupp, arenatyp, log(distans)·rekordtyp, log(distans)·kön, log(distans)·åldersgrupp, log(distans)·arenatyp, rekordtyp·kön, rekordtyp·åldersgrupp, rekordtyp·arenatyp, kön·åldersgrupp, kön·arenatyp och åldersgrupp·arenatyp. De resultat som fås genom att använda de olika procedurerna sammanfattas enligt nedanstående tabell.

Tabell 8

Stegvis regressionsmetod	**Variabler i modellen**	**R^2**
Backward elimination	log(distans), $\log(\text{distans})^2$, $\log(\text{distans})^3$, $\log(\text{distans})^4$, rekordtyp, kön, arenatyp, åldersgrupp, log(distans)·rekordtyp, log(distans)·kön, rekordtyp·kön, rekordtyp·arenatyp	0.9999
Forward selection	log(distans), $\log(\text{distans})^2$, $\log(\text{distans})^3$, $\log(\text{distans})^4$, rekordtyp, kön, arenatyp, åldersgrupp, log(distans)·rekordtyp	0.9999
Stepwise regression	Enligt Forward selection	0.9999

Val av variabler med stegvis regression.

Procedurerna resulterar i två olika modellformuleringar, vilka båda har samma höga förklaringsgrad på 99.99 % som för den ursprungliga modellen innehållandes 18 variabler. Den modell som erhålls av både Forward selection och Stepwise regression innehåller nio variabler, vilket är tre variabler mindre än i den modell som fås enligt Backward elimination. Då modellerna till synes beskriver observationerna lika väl och i övrigt använder samma variabeluppsättning anses den förra enklare modellen vara att föredra. Den enklare modellen har samma förklaringsgrad som den andra, men däremot en något större variansskattning $\hat{\sigma}^2 = 0.0003$. Variablerna har inkluderats i modellen enligt samma ordning för både Forward selection och Stepwise regression, vilket kan ge oss upplysningar om deras relativa betydelse

för regressionsmodellen. Vi tittar därför på stegen i proceduren Stepwise regression enligt tabellen nedan (gäller för typ I-fel).

Tabell 9

Steg	Tillagd variabel	Partiell R^2	Modell R^2	F	p-värde
1	log(distans)	0.9967	0.9967	34711.4	<.0001
2	kön	0.0019	0.9986	149.13	<.0001
3	rekordtyp	0.0004	0.9990	51.78	<.0001
4	$\log(\text{distans})^2$	0.0003	0.9993	46.11	<.0001
5	$\log(\text{distans})^4$	0.0004	0.9997	125.01	<.0001
6	$\log(\text{distans})^3$	0.0001	0.9998	45.94	<.0001
7	åldersgrupp	0.0001	0.9998	38.04	<.0001
8	arenatyp	0.0000	0.9999	22.30	<.0001
9	log(distans) ·rekordtyp	0.0000	0.9999	7.28	0.0081

Resultat av Stepwise regression.

Variablerna har lagts till i den ordning som de visar samband med log(tid). Följaktligen kommer log(distans) med den högsta signifikansnivån att inkluderas i procedurens första steg. Av de återstående variablerna fås sedan kön som mest signifikant, då den förklarar så mycket som ca 58 % av den återstående oförklarade variationen. I steg tre inkluderas rekordtyp som förklarar ca 29 % av den återstående oförklarade variationen. Därefter inkluderas samtliga tre potenser av log(distans) i steg 4-6. Som tidigare nämnts tyder dessa variablers signifikans på att sambandet mellan log(distans) och log(tid) är icke-linjärt, men då potensernas effekter inkluderas först i steg 4-6 graderas relevansen för variablerna som något lägre än för de innan. Proceduren fortsätter sedan att inkludera de två dummyvariablerna åldersgrupp och arenatyp, vilka alltså inkluderas i modellen först vid steg 7-8. Vi vill av den anledningen titta närmare på deras respektive inverkan på log(tid). Först genomförs en multipel linjär regression för en uppsättning variabler enligt steg 1-6 och sedan enligt steg 1-7. Båda modellerna har samma justerade förklaringsgrad på 99.98 % och åldersgrupp fås i den större modellen som starkt signifikant (p-värde <.0001). En multipel linjär regressionsanalys genomförs sedan för en uppsättning variabler enligt steg 1-8, vilket resulterar i en något högre justerad förklaringsgrad på 99.99 %. Även arenatyp fås som starkt signifikant (p-värde <.0001). Jämförelser av modellens parameterskattningar för åldersgrupp respektive arenatyp med de för de övriga dummyvariablerna visar att dessa fås som jämförelsevis små, men på grund av deras starka signifikans och att de anses enkla att beskriva i modellen så väljer vi ändå att inkludera dem. Därtill är vi särskilt intresserade av att undersöka hur de olika kategoriernas effekter ser ut. Sist inkluderas samspelstermen log(distans)·rekordtyp i modellen, vilken får ett jämförelsevis betydligt högre p-värde än de övriga variablerna. Den justerade förklaringsgraden är för modellen oförändrad som 99.99 %. Samspelstermen har ett mindre t-värde än övriga variabler i modellen och likaså en liten parameterskattning. Variabeln log(distans)·rekordtyp har således den minsta effekten på responsvariabeln log(tid) och anses inte tillföra tillräckligt mycket ny information för att det ska vara motiverat att inkludera den i modellen.

Sammantaget väljer vi alltså att utgå från den modell som erhålls med Forward selection och Stepwise regression, men förenklar den något genom att exkludera förklaringsvariabeln log(distans)·rekordtyp. Resultatet blir en något mindre komplicerad modell fri från samspelstermer utan att vi för den delen förlorar för mycket i förklaringskraft. Variablerna i

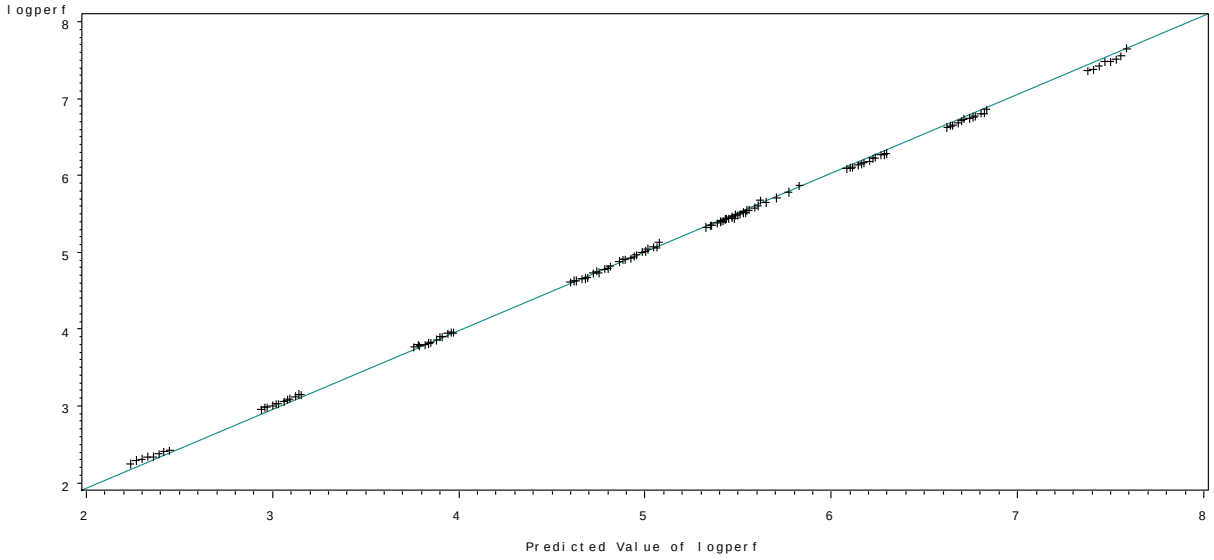
denna regressionsmodell förklarar så mycket som 99.99 % av variationen i log(tid), vilket innebär att de återstående 0.01 % som blir kvar som oförklarad variation därmed kan antas bero på faktorer som inte finns representerade i modellen.

Vi har därmed resonerat oss fram till att den modell som bäst anses kunna beskriva de observerade rekorden har formen

$$log(tid_i) = \log(\alpha) + \beta_1 \cdot log(distans_i) + \beta_2 \cdot log(distans_i)^2 + \beta_3 \cdot log(distans_i)^3 + \beta_4 \cdot log(distans_i)^4 + \beta_5 \cdot rekordtyp + \beta_6 \cdot kön + \beta_7 \cdot åldersgrupp + \beta_8 \cdot arenatyp + \varepsilon_i$$

För att kontrollera modellens lämplighet plottas dess predikterade värden mot de observerade värdena, där observationer som passar perfekt för modellen därmed fås utmed en rak linje med lutningen 1.

Figur 11



Scatterplot: predikterade värden i förhållande till observerade värden av log(tid).

De predikterade och observerade värdena ser ut att stämma bra överrens. Det kan dock observeras att modellen ger en viss underskattning av log(tid) för kortare distanser och en viss överskattning för längre distanser.

Vidare ger en anpassning av denna regressionsmodell med hjälp av minsta-kvadratmetoden resultat enligt nedanstående ANOVA och parameterskattningar.

Tabell 10

Variationskälla	Fg	Kvs	Mkvs	F	R^2/R^2_{adj}
Regression	8	228.6128	28.5766	97329	0.9999/0.9999
Residual	107	0.0314	0.0003		
Total	115	228.64424			

ANOVA: förklaringsvariabler log(distans), log(distans)^2, log(distans)^3, log(distans)^4, rekordtyp, kön, åldersgrupp, arenatyp.

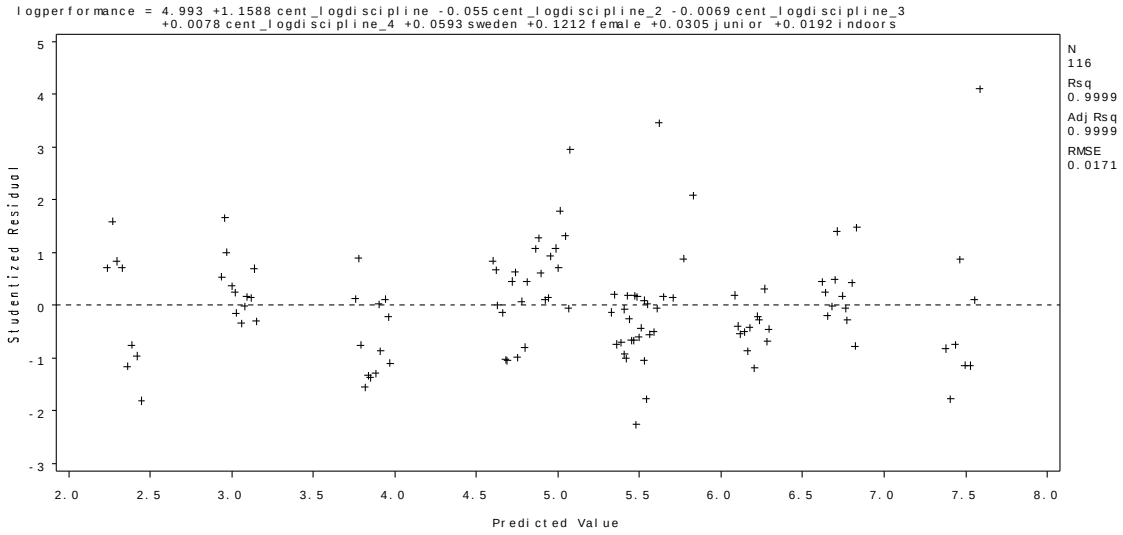
Tabell 11

Variabel	Fg	Parameter-skattning	Medelfel	p-värde	95 % Ki
Intercept	1	4.9930	0.0039	<.0001	(4.985, 5.001)
log(distans)	1	1.1588	0.0030	<.0001	(1.153, 1.165)
log(distans)^2	1	-0.0550	0.0031	<.0001	(-0.061, -0.049)
log(distans)^3	1	-0.0069	0.0008	<.0001	(-0.009, -0.005)
log(distans)^4	1	0.0078	0.0006	<.0001	(0.007, 0.009)
rekordtyp	1	0.0593	0.0032	<.0001	(0.053, 0.066)
kön	1	0.1212	0.0032	<.0001	(0.115, 0.128)
åldersgrupp	1	0.0305	0.0038	<.0001	(0.023, 0.038)
arenatyp	1	0.0192	0.0041	<.0001	(0.011, 0.027)

Parameterskattningar: förklaringsvariabler log(distans), log(distans)^2, log(distans)^3, log(distans)^4, rekordtyp, kön, åldersgrupp, arenatyp.

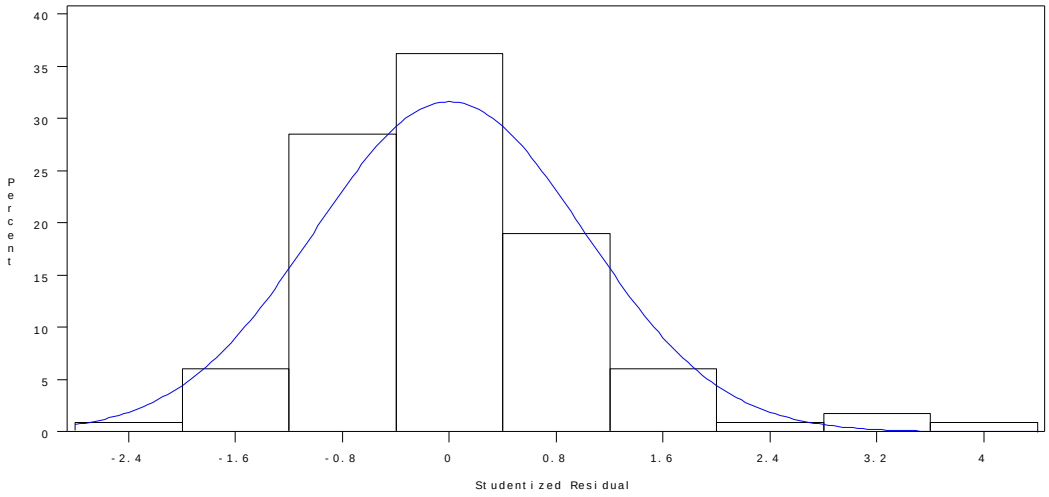
Innan vi accepterar modellen och har förtroende för att korrekta slutsatser kan dras från den behöver vi förvissa oss om att modellförutsättningarna är uppfyllda. Vi vill alltså kontrollera att modellens residualer är oberoende, normalfördelade och har en konstant varians, vilket lämpligen görs med hjälp av residualplots över modellens standardiserade residualer.

Figur 12



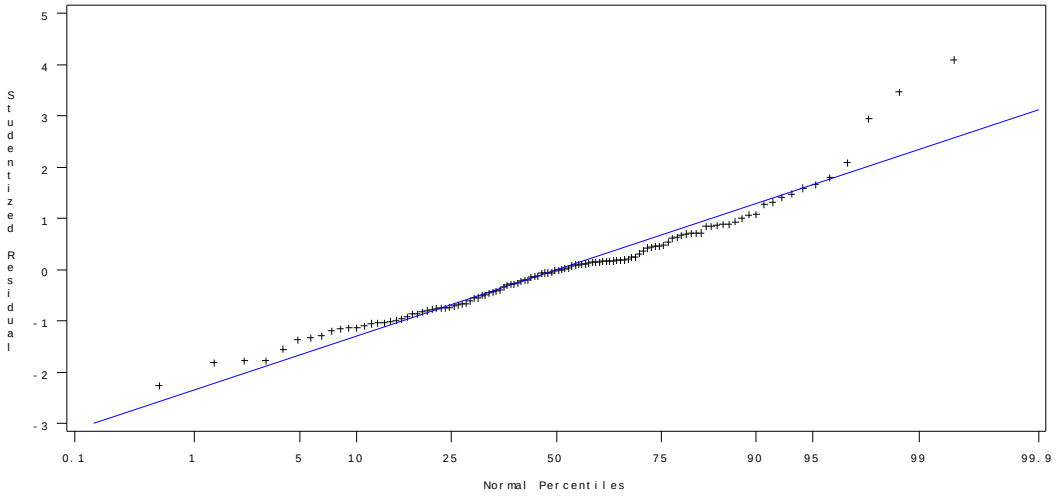
Residualplot: standardiserade residualer i förhållande till predikterade värden.

Figur 13



Residualplot: histogram för standardiserade residualer med normalfördelningskurva.

Figur 14



Residualplot: standardiserade residualer i förhållande till normalpercentilen.

Det går här inte att urskilja något systematiskt mönster och antagandet om residualernas oberoende anses därför vara uppfyllt. Andra och tredje residualploten visar tydligt att det finns en något större spridning i dess ändar och mitt, vilket tyder på en viss avvikelse från normalfördelning. Dessutom observeras ett fåtal avvikande positiva residualer utan några motsvarande stora negativa residualer. Två observationer som visar sig svagare än förväntat tillhör distanserna 1 mile och 10 000 meter. De hör dessutom båda till samma kombination av de fyra olika kategorierna: svenska rekord för kvinnliga juniorer på utomhusarenor. Dessa två observationer stämmer inte riktigt med modellen och skulle eventuellt kunna uteslutas för att ge ett jämnare resultat. Vi väljer dock att behålla dem av den enkla anledningen att vi försöker få en så pass realistisk analys av rekordstatistiken som möjligt. Då linjär regressionsanalys är en relativt robust metod mot mindre avvikelser från antagandet om normalfördelning kan modellförutsättningarna trots dessa mindre avvikelser anses vara approximativt uppfyllda.

4.3 Kovariansanalys

Effekten av de fyra olika kategorivariablerna i datamaterialet kan undersökas närmare genom tillämpning av kovariansanalys, vilket i praktiken innebär att de kategorier som undersöks nu istället beskrivs som klassvariabler. Analysen inleds med en väldigt omfattande modell som inkluderar samtliga variabler som kan tänkas ha en effekt på log(tid). Baserat på variablernas betydelse utesluter vi sedan successivt de som anses onödiga och försöker på så sätt finna en så pass förenklad och lättolkad modell som möjligt. Kovariansanalysen kommer att göras på samma data som i regressionsanalysen i avsnittet ovan, det vill säga det data som har justerats för effekt av reaktionstid, reducerats, log-transformerats och medelvärdescentrerats.

Samtliga 24 variabler som ingår i den ursprungliga modellen, varav 16 är produkttermer, redovisas nedan tillsammans med en sammanfattning av resultatet som fås av en ANCOVA.

Tabell 12

Källa	Typ III SS	F	Pr > F
log(distans)	42.0304	190793	<.0001
log(distans)^2	0.0931	422.68	<.0001
log(distans)^3	0.0220	99.94	<.0001
log(distans)^4	0.0510	231.46	<.0001
rekordtyp	0.0605	274.67	<.0001
kön	0.2609	1184.17	<.0001
åldersgrupp	0.0195	88.69	<.0001
arenatyp	0.0066	29.90	<.0001
rekordtyp·kön	0.0017	7.80	0.0063
rekordtyp·åldersgrupp	0.0006	2.75	0.1006
rekordtyp·arenatyp	0.0007	3.17	0.0781
kön·åldersgrupp	0.0001	0.34	0.5614
kön·arenatyp	0.0000	0.17	0.6786
åldersgrupp·arenatyp	.	.	.
log(distans)·rekordtyp	0.0015	6.90	0.0101
log(distans)·kön	0.0005	2.11	0.1497
log(distans)·åldersgrupp	0.0002	1.03	0.3136
log(distans)·arenatyp	0.0003	1.24	0.2691
log(distans)·rekordtyp·kön	0.0000	0.00	0.9812
log(distans)·rekordtyp·åldersgrupp	0.0015	6.82	0.0105
log(distans)·rekordtyp·arenatyp	0.0000	0.05	0.8189
log(distans)·kön·åldersgrupp	0.0002	0.72	0.3998
log(distans)·kön·arenatyp	0.0002	0.85	0.3584
log(distans)·åldersgrupp·arenatyp	.	.	.

ANCOVA: förklaringsvariabler i ursprunglig modell.

Den anpassade kovariansmodellen har en förklaringsgrad på 99.99 % och en variansskattning på $\widehat{\sigma^2} = 0.0002$ med 93 frihetsgrader. Produkttermerna i tabellen representerar samspel mellan två eller tre olika variabler. Att resultat saknas för samspelet mellan kategorierna åldersgrupp och arenatyp förklaras av att åldersgruppens nivå junior endast har observationer för arenatypens nivå utomhus. Det observeras att en klar majoritet av de effekter som representeras i modellen får låga p-värden, däribland flera av samspelstermerna. Av de icke-signifikanta samspelstermerna fås jämförelsevis låga p-värden för rekordtyp·åldersgrupp, rekordtyp·arenatyp och log(distans)·kön. Det tyder på att det finns ett visst samspel mellan dessa faktorer, vilket innebär att effekten av rekordtyp till viss del beror på löparens tillhörighet i kategorierna åldersgrupp och arenatyp samt att distansens effekt på prestationen är olika för könen.

Det är önskvärt att reducera den väldigt omfattande modellen då flertalet av de inkluderade variablerna inte kan påvisas ha några signifikanta effekter på log(tid). En förenkling görs genom att exkludera de variabler för vilka värden saknas och därefter på 1 % -signifikansnivå successivt exkludera den minst signifikanta variabeln från modellen. Resultatet av detta blir att totalt 11 variabler plockas bort och kvar blir en modell med 13 signifikanta variabler.

Genom att använda kovariansanalys för denna modell fås en kvadratsummeuppdelning enligt nedanstående tabell.

Tabell 13

Källa	Typ III SS	F	Pr > F
log(distans)	42.1649	194496	<.0001
$\log(\text{distans})^2$	0.0931	429.51	<.0001
$\log(\text{distans})^3$	0.0220	101.56	<.0001
$\log(\text{distans})^4$	0.0510	235.20	<.0001
rekordtyp	0.0714	329.48	<.0001
kön	0.4257	1963.73	<.0001
åldersgrupp	0.0195	90.13	<.0001
arenatyp	0.0066	30.38	<.0001
rekordtyp·kön	0.0017	7.93	0.0059
rekordtyp·arenatyp	0.0018	8.10	0.0054
log(distans)·rekordtyp	0.0026	12.14	0.0007
log(distans)·rekordtyp·åldersgrupp	0.0016	7.54	0.0072
log(distans)·kön·arenatyp	0.0021	4.85	0.0097

ANCOVA: förklaringsvariabler i reducerad modell.

Samtliga p-värden är väldigt låga, vilket inträffar efter det att den sista icke-signifikanta variabeln rekordtyp·åldersgrupp (p-värde 0.0943) har exkluderats. Modellen får en hög förklaringsgrad på 99.99 % och har därmed inte påverkats märkbart av att så mycket som 11 variabler har plockats bort från den ursprungliga modellen. Modellen får en variansskattning $\widehat{\sigma^2} = 0.0002$ med 99 frihetsgrader och alltså fås även denna mer eller mindre oförändrad i jämförelse med den för vår ursprungliga modell. Däremot har denna skattning ett större antal frihetsgrader och till följd därav en större tillförlitlighet.

Samtliga termer i den reducerade modellen fås som individuellt statistiskt signifikanta på 1 % -nivån, men eftersom att många effekter testas samtidigt är det mycket troligt att några av dessa egentligen inte är verklig. Vidare anses det tveksamt att samspelstermerna skulle bidra med så pass mycket information till modellen att det skulle motivera att de bör inkluderas. Medeleffekterna av rekordtyp och dess potenser samt de fyra kategorivariablerna är alla av olika storlek, men de anses likväl ha en självklar plats i modellen eftersom att vi är särskilt intresserade av att kunna göra relevanta jämförelser dem emellan. I jämförelse med dessa bidrar övriga variabler som representerar samspelseffekter mellan olika faktorer i modellen med betydligt mindre kvadratsummor och analysen resulterar därmed i relativt höga p-värden. Det kan av den orsaken ifrågasättas vilken vinst som görs av att inkludera dessa fem variabler, särskilt då vi inte har kunnat identifiera någon defekt eller bristande anpassning som skulle motivera att samspelstermerna bidrar med en betydlig förbättring av modellen. Visserligen beskriver en modell utan samspelstermer inte alla de systematiska effekter i datat som det är möjligt att beskriva, men då fördelen med att inkludera dessa anses vara så pass försumbar dras slutsatsen att det inte anses vara motiverat att representera effekterna av samspelstermerna i tabellen ovan.

Vi har därmed resonerat oss fram till en modell på samma form som vid regressionsanalysen i föregående avsnitt. Modellen har således en tillhörande ANOVA-tabell enligt Tabell 10, parameterskattningar enligt Tabell 11 och residualplots enligt Figur 12-14. Tabellen nedan sammanfattar resultatet av en ANCOVA för modellen.

Tabell 14

Källa	Typ III SS	F	Pr > F
log(distans)	43.1210	146866	<.0001
$\log(\text{distans})^2$	0.0927	315.85	<.0001
$\log(\text{distans})^3$	0.0227	77.45	<.0001
$\log(\text{distans})^4$	0.0509	173.34	<.0001
rekordtyp	0.1020	347.44	<.0001
kön	0.4258	1450.27	<.0001
åldersgrupp	0.0194	66.06	<.0001
arenatyp	0.0065	22.30	<.0001

ANCOVA: förklaringsvariabler i slutgiltig modell.

4.4 Ren ANOVA-modell

Genom att log(distans) tillåts vara en kategorivariabel i stället för ett fyragradspolynom enligt tidigare avsnitt fås en modell med ett godtyckligt distansberoende där varje enskilt värde på log(distans) får ett parametervärde. Distans förlorar därmed sin kvantitativa mening och fler parametrar förs in i modellen, men samtidigt fås den då utan några invecklade potens- eller samspelstermer. Vi väljer att benämna denna modell för en ren ANOVA-modell, vilken således beskrivs enligt funktionen

$$log(tid_i) = \log(\alpha) + \beta_j \cdot log\left(distans_j\right) + \beta_{12} \cdot rekordtyp$$
$$+\beta_{13} \cdot kön + \beta_{14} \cdot åldersgrupp + \beta_{15} \cdot arenatyp + \varepsilon_i$$

där j=1,…,11 representerar de olika värdena på log(distans). Då en variansanalys genomförs för modellen fås ett resultat enligt nedanstående ANOVA-tabell.

Tabell 15

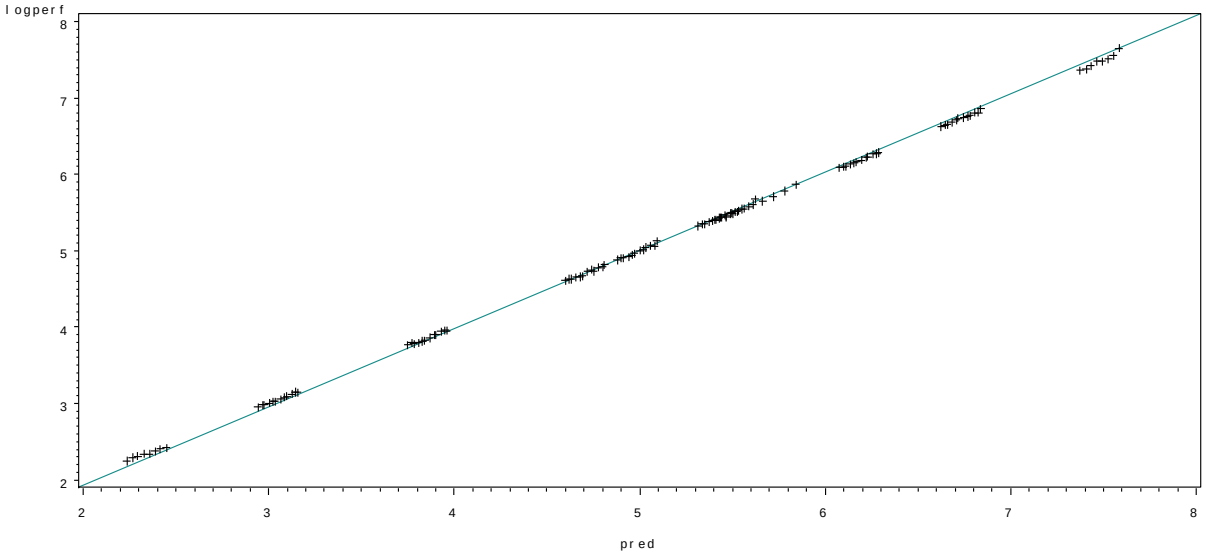
Variationskälla	FG	Kvs	Mkvs	F	R^2
Modell	14	228.6214	16.3301	72163.8	0.9999
Residual	101	0.0229	0.0002		
Total	115	228.64424			

ANOVA: förklaringsvariabler log(distans), rekordtyp, kön, åldersgrupp, arenatyp.

Modellen får en avrundad variansskattning på $\widehat{\sigma^2} = 0.0002$, vilken är något mindre än för den modell vi formulerade utifrån undersökningar med regressions- och kovariansanalys.

Vi kan undersöka den rena ANOVA-modellens lämplighet genom att plotta dess predikterade värden mot observerade värden, där observationer som passar perfekt för modellen fås utmed en rak linje med lutningen 1.

Figur 15



Scatterplot: predikterade värden i förhållande till observerade värden av log(tid).

Modellens värden fås väldigt väl överrensstämmande med de observerade värdena på log(tid). De rekord som passar in bäst på modellen är de tillhörande distansen 400 meter. Samtidigt ger modellen en viss underskattning för de kortare distanserna och en viss överskattning för de längre distanserna. Detta mönster skulle bland annat kunna förklaras av en mindre avvikelse från linjäritet i sambandet, men då punkternas avstånd till linjen är så pass korta anses likväl anpassning vara mycket bra.

Parameterskattningarna för distansvariationen i modellen är inte av väsentligt intresse ur en tolkningssynpunkt, men det är däremot parameterskattningarna för de fyra kategorivariablerna. Dessa fås med tillhörande konfidensintervall enligt nedanstående tabell.

Tabell 16

Variabel	Fg	Parameter-skattning	Medelfel	p-värde	95 % Ki
Intercept	1	7.3749	0.0059	<.0001	(7.3629,7.3864)
rekordtyp	1	0.0593	0.0028	<.0001	(0.0538,0.0649)
kön	1	0.1212	0.0028	<.0001	(0.1156,0.1267)
åldersgrupp	1	0.0318	0.0037	<.0001	(0.0252,0.0385)
arenatyp	1	0.0204	0.0037	<.0001	(0.0132,0.0277)

Parameterskattningar: förklaringsvariabler rekordtyp, kön, åldersgrupp, arenatyp.

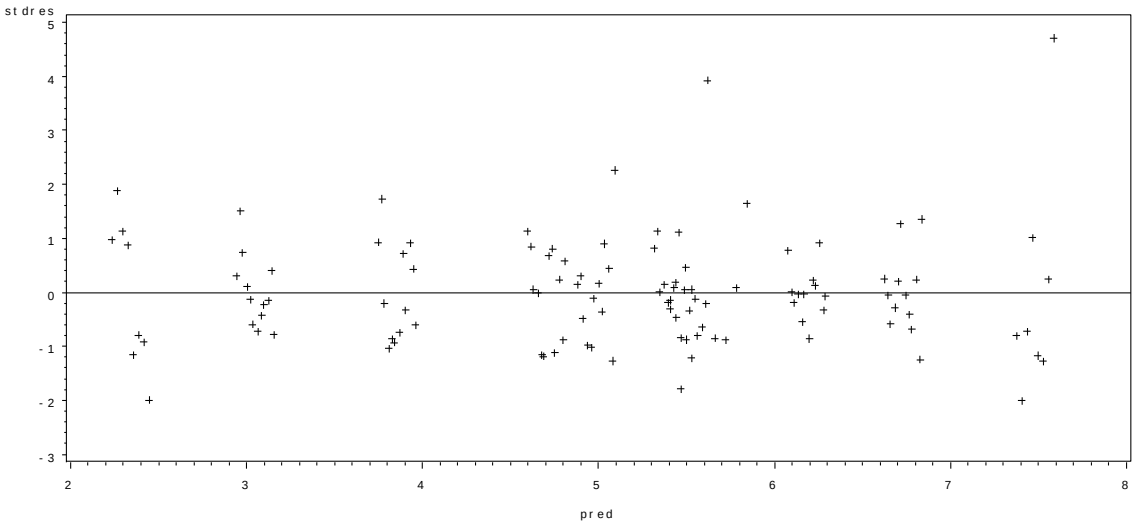
Vidare resulterar modellen i en ANCOVA med tillhörande residualplots enligt nedan. Det kan här poängteras att den variationskälla som vi kallar distansvariation har 10 frihetsgrader till skillnad från de övriga faktorerna som har 1 frihetsgrad.

Tabell 17

Källa	Typ III SS	F	Pr > F
distansvariation	228.0670	100784	<.0001
rekordtyp	0.1020	450.80	<.0001
kön	0.4258	1881.69	<.0001
åldersgrupp	0.0203	89.51	<.0001
arenatyp	0.0070	31.05	<.0001

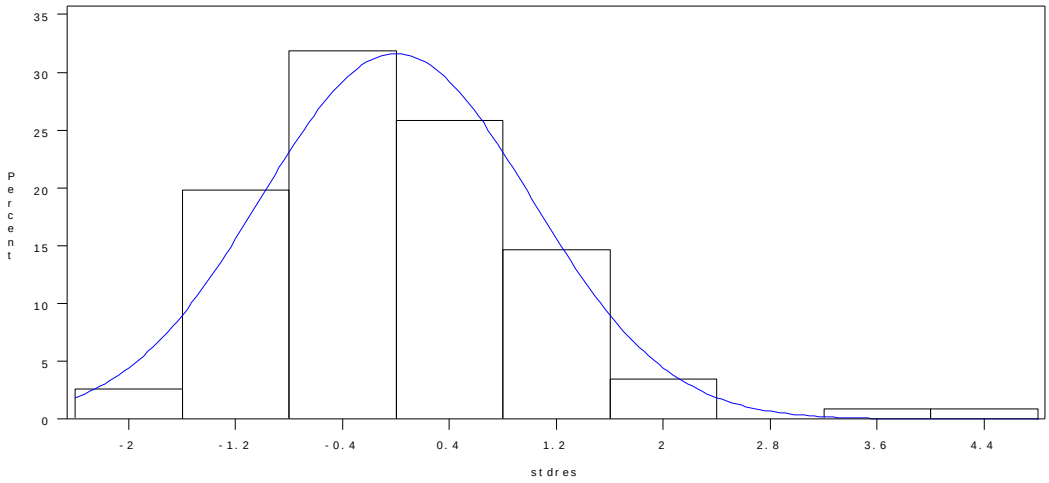
ANCOVA: förklaringsvariabler distansvariation, rekordtyp, kön, åldersgrupp, arenatyp.

Figur 16



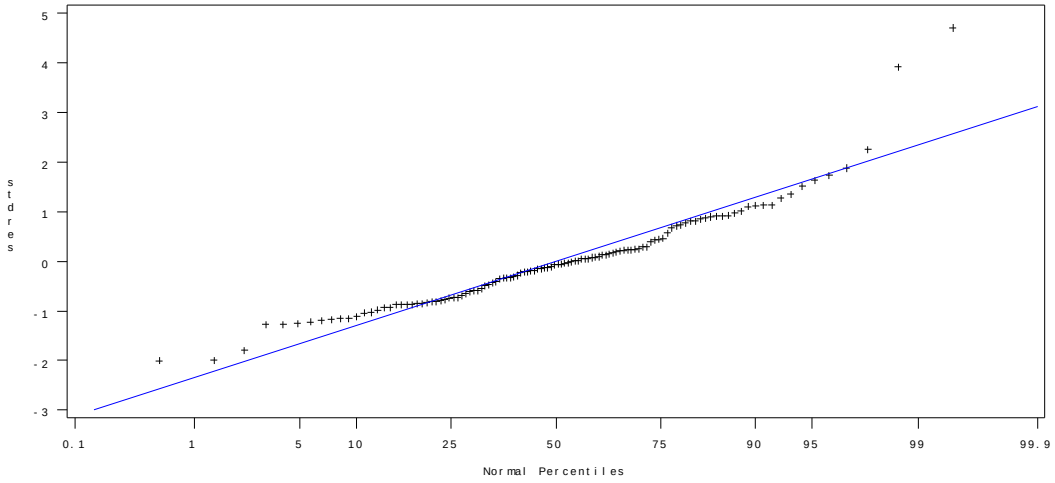
Residualplot: standardiserade residualer i förhållande till predikterade värden.

Figur 17



Residualplot: histogram för standardiserade residualer med normalfördelningskurva.

Figur 18



Residualplot: standardiserade residualer i förhållande till normalpercentilen.

Att döma av modellens residualplots är residualerna någorlunda oberoende av varandra, men precis som vid kontroll av modellantaganden för den slutgiltiga modellen av regressions- och kovariansanalysen ser vi här mindre avvikelser från normalfördelning. Dessa avvikelser anses dock inte heller i detta fall vara av tillräcklig storlek för att det ska ifrågasätta huruvida modellantaganden kan antas vara åtminstone approximativt uppfyllda.

5 Resultat och slutsatser

Den analys som har genomförts för de fem faktorerna distans, rekordtyp, kön, åldersgrupp och junior resulterar i två skilda modellformuleringar, vilka båda anses vara lämpliga för att beskriva variationen i löparnas prestationer. En jämförelse av respektive modells lämplighet fås genom att studera deras sammanfattade egenskaper enligt nedanstående tabell.

Tabell 18

Modell	Antal variabler	Variansskattning	R^2
Linjär regressionsmodell	8	0.0002936	0.9999
Ren ANOVA-modell	14	0.0002263	0.9999

Jämförelse av modeller.

För den linjära regressionsmodellen fås en ANOVA enligt Tabell 10. De parameterskattningar och konfidensintervall som fås enligt Tabell 11 transformeras här tillbaka (anti-logaritmeras) till originalskala för att meningsfulla tolkningar av faktorerna ska kunna göras.

Tabell 19

Variabel	Anti-logaritmerad parameterskattning	Anti-logaritmerat 95 % Ki
Intercept	147.3779	(146.204, 148.562)
log(distans)	3.1861	(3.168, 3.206)
$\log(\text{distans})^2$	0.9465	(0.941, 0.952)
$\log(\text{distans})^3$	0.9931	(0.991, 0.995)
$\log(\text{distans})^4$	1.0078	(1.007, 1.009)
rekordtyp	1.0611	(1.054, 1.068)
kön	1.1289	(1.122, 1.137)
åldersgrupp	1.0310	(1.023, 1.039)
arenatyp	1.0194	(1.011, 1.027)

Parameterskattningar för linjär regressionsmodell i originalskala.

Det skattade interceptet ges här inte någon tolkning av den orsaken att det i datamaterialet inte existerar några rekord nära värdet 0 på distans. Då regressionsmodellen inte innehåller några samspelstermer kan effekten av distans antas vara lika stor för alla de olika kategorinivåerna. Parameterskattningarna för dummyvariablerna representerar följaktligen skillnader mellan nivåerna med avseende på det genometriska medelvärdet. Det observeras att samtliga dessa skattningar uppvisar positiva tecken, vilket innebär att löparna tillhörande de nivåer som representeras av en dummyvariabel med värdet 1 tenderar att prestera jämförelsevis sämre.

Den alternativa modell som vi har valt att benämna ren ANOVA-modell resulterar i en lika hög justerad förklaringsgrad som den linjära regressionsmodellen, men däremot fås en något mindre variansskattning. En ANOVA för modellen fås enligt Tabell 15. Parameterskattningar och konfidensintervall enligt Tabell 16 transformeras även här tillbaka till sin originalskala.

Tabell 20

Variabel	Anti-logaritmerad parameterskattning	Anti-logaritmerat 95 % Ki
Intercept	1595.4323	(1576.402,1613.887)
rekordtyp	1.0611	(1.055,1.067)
kön	1.1289	(1.123,1.135)
åldersgrupp	1.0323	(1.026,1.039)
arenatyp	1.0206	(1.013,1.028)

Parameterskattningar för ren ANOVA-modell i originalskala.

Av samma anledning som tidigare ges här inte det skattade interceptet någon tolkning. Vi ser vid en jämförelse av parameterskattningarna mellan de två olika modellformuleringarna att de i stort överensstämmer med endast marginella skillnader för faktorerna åldersgrupp och arenatyp.

Vi strävar efter att finna en så enkel modell som möjligt som ändå ger en tillräckligt bra beskrivning av det observerade sambandet mellan tid och faktorer. Det är vanskligt att avgöra huruvida en av de erhållna modellerna skulle vara bättre lämpad än den andra, då båda visar jämngoda resultat enligt de kriterier som används vid bedömningen av deras lämplighet. Den rena ANOVA-modellen beskrivs visserligen av betydligt fler variabler, men å andra sidan utan att innehålla några tämligen invecklade samspels- eller potenstermer. Vidare ger denna modell en perfekt anpassning till observationerna. Därmed får den en mindre variansskattning och är i det avseendet, om än litet, så ändå märkbart bättre än den linjära regressionsmodellen. Det finns förvisso inte något formulerat distansberoende i den rena ANOVA-modellen där variabeln log(distans) i princip har förlorat sin kvantitativa mening och till följd av det kan modellen inte användas för prediktering av tider för andra distanser än de som ingår i analysen. Emellertid anses inte heller den linjära regressionsmodellen vara lämpad för det ändamålet eftersom att kännedom saknas om i vilken grad fyragradspolynomet fluktuerar för större värden på distans. Då båda modellerna dessutom resulterar i det närmaste identiska parameterskattningar bedöms det vara motiverat och godtagbart att utgå från den rena ANOVA-modellen för att slutligen kunna dra slutsatser av kategoriernas inverkan på tiden.

Utifrån de erhållna parameterskattningarna för kategorivariablerna i den så kallade rena ANOVA-modellen kan följande slutsatser dras med avseende på det geometriska medelvärdet hos de olika nivåerna givet att övriga variabler hålls konstanta:

Tabell 21

Kategori	Referensnivå	Skattad effekt
rekordtyp	världsrekord	6,1 %
kön	man	12,9 %
åldersgrupp	senior	3,2 %
arenatyp	utomhus	2,1 %

Skattade effekter av kategorivariabler.

Då man bortser från den dominerande inverkan av faktorn distans är det bevisligen löparens kön som har den mest framträdande inverkan på tiden, på så sätt att manliga löpare presterar betydligt bättre än kvinnliga. Det stämmer väl överrens med vad som observerades i inledande undersökningar av datat. Även loppets rekordtyp har en betydande inverkan, då världsrekordsinnehavare presterar märkbart bättre tider jämförelsevis än motsvarande svenska löpare. Vid regressionsanalysen i avsnitt 4.2 kunde en viss samspelseffekt noteras mellan de två variablerna log(distans) och rekordtyp. Detta samspel kan tolkas som att sambandet mellan distans och tid beror på just löparens rekordtyp, det vill säga att distansens effekt är av olika storlek för svenska rekord respektive världsrekord. Den parameterskattning som ges för modellens samspelsterm har positivt tecken, vilket kan tolkas som att världsrekordsinnehavare presterar desto bättre i relation till svenska löpare för växande värden på distans. Emellertid anses samspelseffekten vara så pass marginell att den är umbärlig i en beskrivande modell, men det kan ändå nämnas att en sådan interaktion har noterats mellan dessa två faktorer. Slutligen har även faktorerna åldersgrupp och arenatyp signifikanta effekter på prestationen, men relativt mindre sådana. Seniorer presterar alltigenom bättre än juniorer, vilket anses vara ett rimligt resultat med tanke på deras större träningserfarenhet. Arenatypens inverkan på prestationerna är å andra sidan något mer oförutsägbar. Utifrån det aktuella datamaterialet kan slutsatsen dras att rekordinnehavare generellt presterar 2.1 % bättre tider på utomhusarenor.

6 Diskussion

Det huvudsakliga syftet med detta arbete är att genomföra en statistisk analys av data över rekord i löpning för att bestämma hur olika faktorer inverkar på tiden. Genom tillämpningar av variansanalys för det transformerade datat i logskala erhåller vi två modellvarianter som bedöms vara väl lämpade att beskriva sambandet mellan rekordtid och de studerade faktorerna. Båda dessa modeller använder sig av samtliga tillgängliga faktorer för att beskriva variationen i tiderna, men däremot enligt olika utföranden. Det är svårt att uttala sig om huruvida ena modellen skulle vara överlägsen den andra i något särskilt avseende då de visar jämnbördiga resultat enligt de särskilda kriterier som används för att bedöma deras respektive lämplighet. Kategorieffekterna modelleras på samma sätt i båda modellerna som således bara skiljer sig åt med avseende på hur effekten av variabeln distans representeras. Det har i samtliga undersökta ANOVA- och ANCOVA-tabeller i analysen kunnat noteras att värdena på Typ III SS för kategorivariablerna fås väl överrensstämmande. Med andra ord fås storleken på dessa effekter, efter det att hänsyn först har tagits till modellens övriga variabler, mer eller mindre oförändrad oavsett hur vi väljer att modellera distanseffekten.

Distans visar sig föga förvånande vara den faktor som har den överlägset största effekten på tiden, vilket utgör utgångspunkten för de huvudsakliga undersökningarna av de återstående faktorerna. Vi kan observera signifikanta effekter av samtliga dessa kategorivariabler på löparnas tider. De flesta av skillnaderna mellan nivåerna kan till stor del förutspås redan innan det att analysen genomförs, men det som är av särskilt intresse i denna studie är att se siffror på deras storlek. Den mindre andelen återstående oförklarad variation som vi inte kan förklara med de utvalda modellerna antas till stor del kunna förklaras av individuella effekter, vilka bedöms ha en jämförelsevis mindre betydelse för prestationen. Dessa individuella effekter kan vara av både positiv och negativ karaktär.

Att inga samspelseffekter av betydande storlek kan påvisas i det aktuella datat underlättar analysen av de enskilda kategorivariablernas inverkan. Dessutom kan vi göra tolkningen att effekten av distans är lika stor för båda nivåerna inom en kategori, det vill säga att skillnaden mellan nivåerna är konstant för samtliga värden på distans. Den genomsnittliga tiden skiljer sig alltså åt mellan nivåerna, men förändras i samma hastighet för dem båda.

Vid slutet av avsnitt 4.2 noteras i regressionsmodellens residualplot två oväntat avvikande positiva residualer tillhörande distanserna 1 mile och 10 000 meter. Båda dessa observationer tillhör samma kategorikombination svenska rekord för kvinnliga juniorer på utomhusarenor, vilka följaktligen anses ha en relativt stor förbättringspotential vad gäller rekord på just dessa två distanser. Den bakomvarande orsaken tros vara att det inte har tävlats på dessa distanser i samma omfattning som de övriga. Observationerna stämmer inte något vidare överens med den framtagna regressionsmodellen som ger en särskilt stor underskattning av just dessa tider.

De osäkerhetskällor som främst anses kunna påverka resultaten i den genomförda analysen är osäkerhet på grund av modellantaganden eller bearbetningsfel. I arbetet har en begränsning gjorts till att främst studera olika linjära regressionsmodeller, men det går för den delen inte att utesluta att det finns någon annan typ av funktion som kan användas för att skapa en mer tillfredsställande modell av sambandet. Det är dock oundvikligt att det oavsett val av modell finns en viss grad av osäkerhet kring de erhållna skattningarna. Osäkerheten i skattningarna för de två valda modellerna anses vara tämligen liten baserat på de tillhörande 95 % -iga

konfidensintervallen i Tabell 19-20. Då samtliga parameterskattningars konfidensintervall endast innehåller positiva tal är den skillnaden mellan respektive kategoris nivåer med minst 95 % säkerhet statistiskt säkerställd till fördel för referensnivåerna. För att minimera osäkerhet orsakad av eventuella bearbetningsfel när datat har hämtats på Internet och sedan manuellt matats in i en SAS-tabell har samtliga uppgifter som används i arbetet dubbelkontrollerats. Det går emellertid inte att helt utesluta att det eventuellt kan ha uppstått mindre kodningsfel. Vidare är de modeller som diskuteras här medvetet bristfälliga då vi genom begränsningar väljer att förbise från viss information, till exempel att det rimligtvis bör finnas en viss grad av korrelation mellan de rekordtider som har presterats av en och samma löpare men på olika distanser. Vid en eventuell utvidgning av denna analys skulle man därför kunna undersöka huruvida det finns fler faktorer som har betydelse för prestationerna. Förslag på sådana faktorer är rekordinnehavare, årtal och nationalitet. Ett större dataunderlag skulle samtidigt öka chansen för att lyckas säkerställa små skillnader.

Avslutningsvis bör det framhållas att bedömningarna av de olika modellernas lämplighet till stor del görs utifrån egna tolkningar och preferenser, varför andra bedömningar av resultaten i analysen också skulle kunna leda till att andra slutsatser dras.

7 Referenser

IAAF Athletics (2011). *Records by Category*, http://www.iaaf.org/statistics/recbycat/index.html (Hämtad 2011-08-30)

Milliken, G. A. och Johnson, D. E. (2001). *Analysis of Messy Data, Volume III: Analysis of Covariance*. Chapman & Hall/CRC.

Ohlsson, E (2005). *Kort handledning i SAS*. Matematiska institutionen, Stockholms universitet.

SAS. *SAS Customer Support*, http://support.sas.com/

Sundberg, R. (2010). *Lineära Statistiska Modeller*. Matematiska institutionen, Stockholms universitet.

Svenska friidrottsförbundet. *Allt om grenar – Allmänt*, http://www.friidrott.se/alltom/regler/regler1.aspx (Hämtad 2012-02-17)

Svenska friidrottsförbundet. *Allt om grenar – Löpning*, http://www.friidrott.se/alltom/regler/regler2.aspx (Hämtad 2011-11-07)

Svenska friidrottsförbundet (2010). *Svenska rekord – per 1 januari 2011*, http://www.friidrott.se/rs/rekord/swerek/gallande.aspx (Hämtad 2011-08-30)

Wikipedia (2012). *International Association of Athletics Federations*, http://en.wikipedia.org/wiki/International_Association_of_Athletics_Federations (Hämtad 2012-03-05)

A Appendix

Tabell A1: Svenska rekord, män

Kön	Åldersgrupp	Arenatyp	Distans	Prestation	Rekordinnehavare	Datum
Man	Junior	Utomhus	100	10.47	Edmund Yeboah	2006-07-07
Man	Junior	Utomhus	200	20.75	Johan Engberg	2000-08-07
Man	Junior	Utomhus	400	46.07	Rikard Rasmusson	1992-09-18
Man	Junior	Utomhus	800	107.07	Mattias Claesson	2004-08-07
Man	Junior	Utomhus	1000	144.02	Lars-Åke Joelsson	1979-07-08
Man	Junior	Utomhus	1500	222.07	Morgan Tollofsén	1989-07-24
Man	Junior	Utomhus	1609.34	241.90	Anders Gärderud	1965-08-08
Man	Junior	Utomhus	3000	477.22	Morgan Tollofsén	1988-06-14
Man	Junior	Utomhus	5000	841.66	Gustav Svedbrant	2000-06-03
Man	Junior	Utomhus	10 000	1772.20	Mustafa Mohamed	1988-06-14
Man	Senior	Utomhus	100	10.18	Peter Karlsson	1996-06-09
Man	Senior	Utomhus	200	20.30	Johan Wissman	2007-09-23
Man	Senior	Utomhus	400	44.56	Johan Wissman	2007-08-29
Man	Senior	Utomhus	800	105.45	Rizak Dirshe	2003-07-19
Man	Senior	Utomhus	1000	137.80	Dan Waern	1959-08-21
Man	Senior	Utomhus	1500	216.49	Johnny Kroon	1985-06-27
Man	Senior	Utomhus	1609.34	234.45	Anders Gärderud	1975-06-30
Man	Senior	Utomhus	2000	302.09	Anders Gärderud	1975-07-04
Man	Senior	Utomhus	3000	462.24	Dan Glans	1979-07-05
Man	Senior	Utomhus	5000	797.59	Anders Gärderud	1976-07-05
Man	Senior	Utomhus	10000	1675.74	Jonny Danielson	1989-06-07
Man	Senior	Inomhus	50	5.83	Stefan Nilsson	1981-02-21
Man	Senior	Inomhus	60	6.58	Peter Karlsson	1996-02-07
Man	Senior	Inomhus	200	20.65	Johan Wissman	2004-02-28
Man	Senior	Inomhus	400	45.59	Jimisola Laursen	2002-03-03
Man	Senior	Inomhus	800	105.91	Martin Enholm	1992-02-22
Man	Senior	Inomhus	1000	140.56	Torbjörn Johansson	1996-02-25
Man	Senior	Inomhus	1500	219.92	Jörgen Zaki	1996-02-11
Man	Senior	Inomhus	1609.34	242.30	Ulf Högberg	1968-01-13
Man	Senior	Inomhus	3000	468.44	Kent Claesson	1997-02-14
Man	Senior	Inomhus	5000	819.71	Mustafa Mohamed	2006-07-14

Tabell A2: Svenska rekord, kvinnor

Kön	Åldersgrupp	Arenatyp	Distans	Prestation	Rekordinnehavare	Datum
Kvinna	Junior	Utomhus	100	11.35	Linda Haglund	1975-07-12
Kvinna	Junior	Utomhus	200	23.35	Jenny Ljunggren	2003-07-27
Kvinna	Junior	Utomhus	400	52.23	Ann-Louise Skoglund	1981-08-13
Kvinna	Junior	Utomhus	800	123.90	Lena Nilsson	1998-06-26
Kvinna	Junior	Utomhus	1000	168.09	Malin Ewerlöf	1990-06-29
Kvinna	Junior	Utomhus	1500	247.47	Inger Knutsson	1973-08-26
Kvinna	Junior	Utomhus	1609.34	292.20	Katarina Wåhlin	1983-06-24
Kvinna	Junior	Utomhus	3000	538.36	Inger Knutsson	1973-09-01
Kvinna	Junior	Utomhus	5000	951.17	Jessica Carlberg	1998-08-02
Kvinna	Junior	Utomhus	10 000	2105.30	Charlotte Sinclair	2007-08-10
Kvinna	Senior	Utomhus	100	11.16	Linda Haglun	1980-07-26
Kvinna	Senior	Utomhus	200	22.82	Linda Haglun	1979-07-01
Kvinna	Senior	Utomhus	400	51.69	Ann-Louise Skoglund	1986-07-01
Kvinna	Senior	Utomhus	800	119.44	Malin Ewerlöf	1998-08-19
Kvinna	Senior	Utomhus	1000	158.70	Maria Akraka	1994-08-24
Kvinna	Senior	Utomhus	1500	245.49	Malin Ewerlöf	1997-06-24
Kvinna	Senior	Utomhus	1609.34	265.34	Malin Ewerlöf	1997-08-18
Kvinna	Senior	Utomhus	2000	352.22	Sara Wedlund	1995-09-02
Kvinna	Senior	Utomhus	3000	528.87	Sara Wedlund	1996-06-29
Kvinna	Senior	Utomhus	5000	906.90	Sara Wedlund	1996-07-08
Kvinna	Senior	Utomhus	10000	1917.15	Midde Hamrin	1990-08-19
Kvinna	Senior	Inomhus	50	6.17	Linda Haglund	1981-02-22
Kvinna	Senior	Inomhus	60	7.13	Linda Haglund	1978-03-12
Kvinna	Senior	Inomhus	200	23.47	Maria Staafgård	1994-03-12
Kvinna	Senior	Inomhus	400	52.40	Ann-Louise Skoglund	1986-02-23
Kvinna	Senior	Inomhus	800	120.01	Maria Akraka	1998-02-19
Kvinna	Senior	Inomhus	1000	158.11	Malin Ewerlöf	1999-02-25
Kvinna	Senior	Inomhus	1500	247.74	Maria Akraka	1992-02-18
Kvinna	Senior	Inomhus	1609.34	272.49	Johanna Nilsson	2003-03-15
Kvinna	Senior	Inomhus	3000	530.32	Sara Wedlund	1996-03-09
Kvinna	Senior	Inomhus	5000	906.49	Sara Wedlund	1996-02-25

Tabell A3: Världsrekord, män

Kön	Åldersgrupp	Arenatyp	Distans	Prestation	Rekordinnehavare	Datum
Man	Junior	Utomhus	100	10.01	Darrel Brown	2003-08-24
Man	Junior	Utomhus	200	19.93	Usain Bolt	2004-04-11
Man	Junior	Utomhus	400	43.87	Steve Lewis	1988-09-28
Man	Junior	Utomhus	800	102.69	Abubaker Kaki	2008-06-06
Man	Junior	Utomhus	1000	135.00	Benjamin Kipkurui	1999-07-17
Man	Junior	Utomhus	1500	210.24	Cornelius Chirchir	2002-07-19
Man	Junior	Utomhus	1609.34	229.29	Ilham Tanui Özbilen	2009-07-03
Man	Junior	Utomhus	3000	448.78	Augustine Kiprono Choge	2005-05-13
Man	Junior	Utomhus	5000	772.61	Eliud Kipchoge	2003-06-27
Man	Junior	Utomhus	10 000	1601.75	Samuel Kamau Wanjiru	2005-08-26
Man	Senior	Utomhus	100	9.58	Usain Bolt	2009-08-16
Man	Senior	Utomhus	200	19.19	Usain Bolt	2009-08-20
Man	Senior	Utomhus	400	43.18	Michael Johnson	1999-08-26
Man	Senior	Utomhus	800	101.01	David Lekuta Rudisha	2010-08-29
Man	Senior	Utomhus	1000	131.96	Noah Ngeny	1999-09-05
Man	Senior	Utomhus	1500	206.00	Hicham El Guerrouj	1998-07-14
Man	Senior	Utomhus	1609.34	223.13	Hicham El Guerrouj	1999-07-07
Man	Senior	Utomhus	2000	284.79	Hicham El Guerrouj	1999-09-07
Man	Senior	Utomhus	3000	440.67	Daniel Komen	1996-09-01
Man	Senior	Utomhus	5000	757.35	Kenenisa Bekele	2004-05-31
Man	Senior	Utomhus	10000	1577.53	Kenenisa Bekele	2005-08-26
Man	Senior	Inomhus	50	5.56	Donovan Bailey	1996-02-09
Man	Senior	Inomhus	60	6.39	Maurice Greene	1998-02-03
Man	Senior	Inomhus	200	19.92	Frank Fredericks	1996-02-18
Man	Senior	Inomhus	400	44.57	Kerron Clement	2005-03-12
Man	Senior	Inomhus	800	102.67	Wilson Kipketer	1997-03-09
Man	Senior	Inomhus	1000	134.96	Wilson Kipketer	2000-02-20
Man	Senior	Inomhus	1500	211.18	Hicham El Guerrouj	1997-02-02
Man	Senior	Inomhus	1609.34	228.45	Hicham El Guerrouj	1997-02-12
Man	Senior	Inomhus	3000	444.90	Daniel Komen	1998-02-06
Man	Senior	Inomhus	5000	769.60	Kenenisa Bekele	2004-02-20

Tabell A4: Världsrekord, kvinnor

Kön	Åldersgrupp	Arenatyp	Distans	Prestation	Rekordinnehavare	Datum
Kvinna	Junior	Utomhus	100	10.88	Marlies Göhr	1977-07-01
Kvinna	Junior	Utomhus	200	22.18	Allyson Felix	2004-08-25
Kvinna	Junior	Utomhus	400	49.42	Grit Breuer	1991-08-27
Kvinna	Junior	Utomhus	800	114.01	Pamela Jelimo	2008-08-29
Kvinna	Junior	Utomhus	1000	155.40	Katrin Wühn	1984-07-12
Kvinna	Junior	Utomhus	1500	231.34	Yinglai Lang	1997-10-18
Kvinna	Junior	Utomhus	1609.34	257.57	Zola Pieterse	1985-08-21
Kvinna	Junior	Utomhus	3000	508.83	Zola Pieterse	1985-09-07
Kvinna	Junior	Utomhus	5000	870.88	Tirunesh Dibaba	2004-06-11
Kvinna	Junior	Utomhus	10 000	1826.50	Linet Chepkwemoi Masai	2008-08-15
Kvinna	Senior	Utomhus	100	10.49	Florence Griffith-Joyner	1988-07-16
Kvinna	Senior	Utomhus	200	21.34	Florence Griffith-Joyner	1988-09-29
Kvinna	Senior	Utomhus	400	47.60	Marita Koch	1985-10-06
Kvinna	Senior	Utomhus	800	113.28	Jarmila Kratochvílová	1983-07-26
Kvinna	Senior	Utomhus	1000	148.98	Svetlana Masterkova	1996-08-23
Kvinna	Senior	Utomhus	1500	230.46	Yunxia Qu	1993-09-11
Kvinna	Senior	Utomhus	1609.34	252.56	Svetlana Masterkova	1996-08-14
Kvinna	Senior	Utomhus	2000	325.36	Sonia O´Sullivan	1994-07-08
Kvinna	Senior	Utomhus	3000	486.11	Junxia Wang	1993-09-13
Kvinna	Senior	Utomhus	5000	851.15	Tirunesh Dibaba	2008-06-06
Kvinna	Senior	Utomhus	10000	1771.78	Junxia Wang	1993-09-08
Kvinna	Senior	Inomhus	50	5.96	Irina Privalova	1995-02-09
Kvinna	Senior	Inomhus	60	6.92	Irina Privalova	1993-02-11
Kvinna	Senior	Inomhus	200	21.87	Merlene Ottey	1993-02-13
Kvinna	Senior	Inomhus	400	49.59	Jarmila Kratochvílová	1982-03-07
Kvinna	Senior	Inomhus	800	115.82	Jolanda Batageli	2002-03-03
Kvinna	Senior	Inomhus	1000	150.94	Maria de Lurdes Mutola	1999-02-25
Kvinna	Senior	Inomhus	1500	238.28	Elena Soboleva	2006-02-18
Kvinna	Senior	Inomhus	1609.34	257.14	Doina Melinte	1990-02-09
Kvinna	Senior	Inomhus	3000	503.72	Meseret Defar	2007-02-03
Kvinna	Senior	Inomhus	5000	864.37	Meseret Defar	2009-02-18