## Kunskapskontroll R-programmering

Enligt statistik från SCB ökar antalet bilar i trafik i Sverige årligen. Detta indikerar att ett verktyg som kan användas för att prediktera och förstå vad som påverkar bilpriser kan vara användbart.

Syftet med modelleringen är att dels bygga en modell som kan prediktera försäljningspriser för begagnade bilar, dels att undersöka vilka bilegenskaper som påverkar försäljningspriset mest.

Data för träning och test har samlats in från Blocket.se genom att hämta information från annonser. Det är inte säkert att slutligt försäljningspris faktiskt stämmer överens med det pris som annonserats. Priser har filtrerats när annonser valts till att ligga mellan 150 000 – 500 000 SEK. Datasetet innehåller nio variabler: Pris, Bränsle, Modellår, Växellåda, Miltal, Biltyp, Drivning, Hästkrafter och Märke.

Inledande dataanalys gav flera insikter inför modelleringen och motiverade åtgärder. En första initial modell valdes till linjär regression för att möjliggöra tester och analysera för modellutveckling. Denna modell är enkel att tolka, vilket passar bra när inferens ska göras. Analysen visade på viss skevhet och tendenser på olinjäritet i data, korrelation mellan variabler och att det fanns outliers bland annat.

Log-transformation av priset ledde till mer normalfördelade residualer samt bättre prediktionsförmåga. Residualerna visade dock fortsatt på systematiskt fel. Transformation av obereroende variabler förbättrade inte detta och inkluderades inte i fortsatt modellering. Det systematiska felet beror med stor sannolikhet på att priser har filtrerats när annonser valts på Blocket och därför behövs bättre data för att skapa hantera detta.

Att rensa bort outliers ledde förvisso till en mer stabil modell, men analys av dessa punkter visade att de var legitima. Slutsatsen gjordes att behålla dessa för vidare modellering för att undvika överanpassning till träningsdatan.

Variabeln biltyp var efter korrelationsanalys motiverad att exkludera. Detta ledde dock till en sämre prediktionsförmåga och därför valdes det att låta denna vara kvar.

Fyra modeller valdes för modelleringen: Linjär regression, Linjär regression med best subset selection, Lasso och Ridge. Att använda modeller med regularisering motiverades av analysen och att korrelation fanns indikerat. Lasso var den modell med lägst RMSE och högst R2 och utväderades till den bästa modellen. Detta indikerar att inte alla variabler bidrar till att prediktera priset.

Modellår och hästkrafter är de variabler som analysen visar har störst påverkan på priset, en nyare bil med ett högre antal hästkrafter är dyrare. Vidare kan det även konstateras att till exempel diesel- och elbilar är dyrare än en bensinbil och att det är billigare med en manuell växellåda jämfört med en automat.

Nästa steg för att bygga en än bättre modell vore att samla in mer data som inte är filtrerad och möjligtvis utöka antalet variabler, tex när annonsen publicerades verkar motiverat att inkludera. Att även testa modeller som kan hantera olinjär data kan leda till förbättringar.