Stockholms universitet

Regressionsanalys av priser på mobiltelefoner

Alice Vijitraphongsa

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:9
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2018:9**
http://www.math.su.se/matstat

Regressionsanalys av priser på mobiltelefoner

Alice Vijitraphongsa*

Juni 2018

Sammanfattning

Smarta mobiltelefoner är idag en av de viktigaste hjälpmedlen i vardagslivet. Dock är det inte något som är gratis och inte särskilt billigt heller. Därför kan det vara intressant att undersöka om det finns något statistiskt samband mellan priset och mobiltelefonens specifikationer, vilket är det vi gör i detta arbete. Vi ska även se om det går att hitta en modell som kan prediktera priser. Vi kommer endast ha mobiltelefoner med operativsystemet Android i datamaterialet. Med hjälp av multipel linjär regression och transformationer av variabler kommer vi att komma fram till två olika modeller som uppfyller förutsättningarna för linjär regression. Stegvis variabelselektion används för att selektera bort icke-signifikanta parametrar. Modellen som passar data bäst och som även är bäst lämpad för prediktion har logaritmerad responsvariabel och innehåller förklarande variabler som står för mobiltelefonens skärmupplösning per skärmstorlek (logaritmerad), vikt, interna minne, tillverkare, vattentålighet samt när mobiltelefonen lanserades. Av samtliga variabler som togs med visade det sig att logaritmerad skärmupplösning per skärmstorlek hade störst effekt på priset och att vikt hade minst effekt.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: alice.vijitraphongsa@gmail.com. Handledare: Kristofer Lindensjö & Felix Wahl.

Förord

Arbetet omfattar 15 högskolepoäng och är ett kandidatarbete i matematisk statistik, utfört på matematiska institutionen på Stockholms Universitet. Jag vill tacka mina handledare Felix Wahl och Kristofer Lindensjö för deras hjälp och värdefulla råd. Jag vill även tacka min familj för deras stöd genom hela utbildningen.

Innehåll

1 Inledning **4**
1.1 Frågeställningar 4

2 Teori **4**
2.1 Linjär regression 4
2.1.1 Minsta-kvadratmetoden och parameterskattning . . . 5
2.1.2 Hypotestest och p-värde 6
2.2 Modellval 7
2.2.1 Förklaringsgrad & justerad förklaringsgrad 7
2.2.2 Stegvis variabelselektion 8
2.2.3 Multikollinjäritet 8
2.2.4 Variansinflationsfaktor - VIF 9
2.2.5 Akaikes informationskriterium - AIC 9
2.2.6 Residualer 9
2.3 Transformering av variabler 10
2.3.1 Dummyvariabel 10
2.3.2 Logtransformering 10
2.3.3 Saknade värden 11
2.4 Prediktion 11
2.4.1 Mean squared error of prediction - MSEP 11

3 Datamaterial **12**
3.1 Beskrivning av variabler 12
3.2 Transformationer och behandling av variabler 13
3.3 Slutgiltiga variabler 18

4 Statistisk modellering **19**
4.1 Enkel linjär regression 19
4.2 Multipel linjär regression 20
4.2.1 Stegvis variabelselektion 24
4.3 Modelljämförelse 24
4.4 Prediktion 25

5 Resultat **26**
5.1 Reducerad modell 1 27
5.2 Reducerad modell 3 28

6 Diskussion **30**

A Datamaterial **33**

B Multikollinjäritet **34**

1 Inledning

Användning av smarta telefoner hos både unga och äldre har ökat markant de senaste åren. Nästan 98 procent av befolkningen har en egen mobiltelefon och 81 procent av befolkningen har en smart mobiltelefon [3]. Med dagens smarta telefoner kan man göra allt från att ringa samtal till att titta på film. Priset på mobiltelefoner kan skilja sig väldigt mycket och syftet med denna uppsats är främst att undersöka vilka faktorer som förklarar priset.

För att få svar på detta kommer vi att använda oss av regressionsanalys. Transformationer av variabler och variabelselektioner görs för att hitta signifikanta variabler och slutligen kommer vi att testa modellens prediktionsförmåga. Vi kommer först att gå igenom teorin bakom de statistiska metoderna som kommer att användas och presentera datamängden innan vi börjar med regressionsanalysen. Analysen kommer att baseras på 200 slumpmässigt valda mobiltelefoner med operativsystemet Android som finns på prisjakts hemsida.

1.1 Frågeställningar

I detta arbete kommer vi att besvara följande frågor med hjälp av de statistiska metoderna som gås igenom i avsnitt 2:

- Finns det något statistiskt samband mellan de förklarande variablerna och responsvariabeln?
- Går det att hitta en modell som kan prediktera priset på en mobiltelefon?

2 Teori

I detta avsnitt kommer vi att gå igenom teorin bakom de statistiska metoderna som används i detta arbete för att besvara frågeställningarna.

2.1 Linjär regression

Linjära regressionsmodeller används för att påvisa om en responsvariabel kan förklaras linjärt av en eller flera andra förklaringsvariabler. Följande förutsättningar måste uppfyllas för linjär regression:

> **Linjäritet**: Responsvariabeln och de förklarande variablerna har ett linjärt samband, då formeln för linjär regression är linjär. Om de inte har ett linjärt samband så är bland annat parameterskattningarna inte väntevärdesriktiga, det vill säga väntevärdet för parameterskattningen är inte lika med värdet på den sökta parametern (Andersson & Tyrcha, 2015, s. 59-60).

Normalfördelad: Modellens residualer är normalfördelade med väntevärde 0 (Andersson & Tyrcha, 2015, s. 25). Detta antagande behövs framförallt vid hypotestester.

Homoskedasticitet: Residualerna har konstant varians σ^2. Om residualerna inte har konstant varians är bland annat t-test inte lämpligt (Andersson & Tyrcha, 2015, s. 66). Man kan dock transformera respons- och/eller förklaringsvariabler i försök att få konstant varians.

Ej perfekt multikollinjär: De förklarande variablerna är inte linjärt beroende. Då förklarande variabler har hög multikollinjäritet blir parameterskattningarna missvisande (Andersson & Tyrcha, 2015, s. 65).

I det fall då det endast finns en förklarande variabel så används enkel linjär regression, som definieras som följande

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i\text{=1,...,n}$$

där ϵ_i är slumpfel som är oberoende och normalfördelad med väntevärde 0 och varians σ^2 (Sundberg, 2014, s. 50-51). Om det finns flera förklarande variabler använder vi istället multipel linjär regression, som ser ut som följande

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \epsilon_i, \quad i\text{=1,...,n} \tag{2.1}$$

Formeln (2.1) kan även skrivas på matrisform och ser då ut som följande

$$y = X\beta + \epsilon$$

där $y = (y_1 \ldots y_n)^T$, $\beta = (\alpha, \beta_1 \ldots \beta_m)^T$, $\epsilon = (\epsilon_1 \ldots \epsilon_n)^T$ och

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{mn} \end{bmatrix}$$

där kolumnen med bara ettor är specifierad för interceptet α (Andersson & Tyrcha, 2015, s. 19-20).

2.1.1 Minsta-kvadratmetoden och parameterskattning

Då β är okänd behöver vi skatta den ur data. Vi använder oss av minsta-kvadratmetoden för att skatta β. Med minsta-kvadratmetoden minimerar

vi, med avseende på β, summan av det kvadratiska avståndet mellan observationerna och regressionslinjen, det vill säga $\sum\limits_{i=1}^{n}(y_i - \alpha - \sum_{j=1}^{m} x_{ji}\beta_j)^2$. Minsta-kvadratskattningen för β ges av (Sundberg, 2014, s. 20)

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Då antagandet om att residualerna är normalfördelade med väntevärde 0 och konstant varians σ^2 är uppfyllt kan man visa att (Sundberg, 2014, s. 22)

$$E(\hat{\beta}) = \beta \tag{2.2}$$

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{2.3}$$

Då σ^2 är okänd skattar vi den med datan vi har. En väntevärdesriktig skattning av σ^2 ges av

$$\hat{\sigma}^2 = \frac{1}{n-m} \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{m} x_{ji}\hat{\beta}_j \right)^2$$

där $n - m$ är antalet frihetsgrader och m är antalet β-parametrar i modellen (Sundberg, 2014, s. 25). Med frihetsgrader menas antalet oberoende observationer som kan användas för skattning.

2.1.2 Hypotestest och p-värde

Vi ska testa hypotesen om regressionskoefficienten β_j är signifikant skild från noll i syfte om att undersöka om förklaringsvariabel x_j har någon effekt på responsvariabeln. Det vill säga, vi ska alltså testa nollhypotesen

$$H_0 : \beta_j = 0$$

mot alternativa hypotesen

$$H_a : \beta_j \neq 0.$$

Om antagandet om att residualerna är normalfördelade med väntevärde 0 och konstant varians σ^2 är uppfyllt, följer det av (2.2) och (2.3) att $\hat{\beta}_j$ är normalfördelad med väntevärde β_j och varians $(\sigma^2 (X^T X)^{-1})_{jj}$, där $(\sigma^2 (X^T X)^{-1})_{jj}$ är det j-te diagonalelementet i matrisen (2.3) (Andersson & Tyrcha, 2015, s. 44). Eftersom att vi ska jämföra om två medelvärden har en signifikant skillnad använder vi oss av t-test för att testa hypotesen. Teststatistikan ges av

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{((X^T X)^{-1})_{jj}}}$$

och är t-fördelad med frihetsgrader $(n-m)$ under H_0, där m är antalet β-parametrar i modellen. Vi förkastar nollhypotesen på signifikansnivån α om $|T_j| \geq t_{\alpha/2}(n-m)$, där $t_{\alpha/2}(n-m)$ är den kritiska gränsen och fås ur tabell till t-fördelningens kvantiler. I vårt fall använder vi signifikansnivån $\alpha = 0.05$.

Vi kommer att ta hjälp av p-värdet för att avgöra vilka variabler som ska tas med i modellen. P-värdet definieras som sannolikheten att $|T_j| \geq t_{\alpha/2}(n-m)$ under nollhypotesen, det vill säga $P(|T_j| \geq t_{\alpha/2}(n-m))$. Om p-värdet för en variabel är mindre än $\alpha = 0.05$ så är koefficienten β_j för variabeln x_j signifikant skild från noll och nollhypotesen kan då förkastas. Variabeln x_j tas med i modellen då nollhypotesen kan förkastas (Andersson & Tyrcha, 2015, s. 49)

2.2 Modellval

För att komma fram till en lämplig modell som passar vår data så använder vi oss av olika metoder och anpassningsmått. Dessa förklaras i detta avsnitt.

2.2.1 Förklaringsgrad & justerad förklaringsgrad

Förklaringsgraden, R^2, definieras som andelen av den totala variationen som modellen förklarar (Sundberg, 2014, s. 69) och har formeln

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

Låt $\hat{y}_i$ vara det skattade värdet av y_i och $\bar{y}$ medelvärdet av y, då är Kvs(residual) $= \sum(y_i - \hat{y}_i)^2$ residualkvadratsumman och Kvs(totalt) $= \sum(y_i - \bar{y})^2$ den totala kvadratsumman, det vill säga den totala variationen hos y-värdena.

R^2 antar bara värden mellan 0 och 1 och ju högre värde, desto bättre anpassning har modellen till observationerna. R^2 används mest till att jämföra olika modeller i samma datamängd, då man kan se hur R^2 varierar vid borttagning eller tillägg av en variabel. En nackdel med detta mått är att förklaringsgraden alltid höjs när man lägger till en variabel, även om variabeln inte förklarar responsvariabeln. I sådant fall är det bättre att använda måttet *Adjusted* R^2, som mäter hur mycket variationen minskats i den aktuella modellen och tar hänsyn till antalet parametrar. *Adjusted* R^2 definieras som följande

$$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2},$$

där $\hat{\sigma}_0^2 = \text{Kvs(totalt)}/(n-1)$ är σ^2-skattningen när det inte finns någon x-variabel i modellen (Sundberg, 2014, s. 70).

2.2.2 Stegvis variabelselektion

När man har en stor uppsättning potentiella förklaringsvariabler är det sällan att alla variabler behövs i modellen. Istället för att undersöka alla möjliga kombinationer av variabler så kan vi använda oss av stegvis variabelselektion, som inte kräver lika mycket beräkningsarbete. Det innebär att vi successivt inkluderar eller eliminerar en variabel i taget tills något stoppkriterium är uppfyllt. I analysen av mobiltelefonpriserna kommer vi att använda oss av följande tre procedurer i stegvis variabelslektion (Sundberg, 2014, s. 71-72):

> **Backward elimination**: Proceduren utgår från modellen som innehåller samtliga variabler. Parametrar som inte är signifikant skilda från noll elimineras successivt tills alla återstående parametrar är signifikanta. Om flera variabler är icke-signifikanta elimineras den variabeln som är mest icke-signifikant, det vill säga variabeln som har högst p-värde.
>
> **Forward selection**: Denna procedur utgår från en modell med endast interceptet. Variabler läggs till en åt gången. Vilken variabel som läggs till bestäms av hur lågt p-värde den har. Variabler läggs till tills det inte längre finns återstående signifikanta variabler att lägga till.
>
> **Stepwise regression**: Denna procedur är en blandning av backward elimination och forward selection. Likt forward selection så utgår proceduren från en modell utan några variabler och lägger till en variabel i taget. Skillnaden är att i stepwise regression så elimineras eventuella variabler som blir icke-signifikanta vid tillägg av en ny variabel.

Dessa procedurer kan dock ge missvisande resultat om antalet variabler är för stort eller om variablerna är korrelerade.

2.2.3 Multikollinjäritet

När man har flera potentiella förklaringsvariabler finns det risk för att två eller fler förklaringsvariabler har approximativt linjära samband. Då säger man att det förekommer multikollinjäritet bland variablerna. Detta medför komplikationer då till exempel två förklarande variabler som enskilt har ett signifikant samband med responsvariabeln kan visa sig vara icke-signifikanta när båda variablerna ingår i modellen. Detta går att åtgärdas genom att till exempel ta bort en av förklaringsvariablerna som utgör det approximativt

linjära sambandet. Genom att undersöka VIF-värderna (se avsnitt 2.2.4) i modellen kan man se om det råder multikollinjäritet.

2.2.4 Variansinflationsfaktor - VIF

Variansinflationsfaktorn, VIF, är ett statistiskt mått på multikollinjäritet. Vi har att variansen av den j-te koefficienten är

$$\mathrm{Var}(\hat{\beta}_j) = \left(\frac{\sigma^2}{\sum_{i=1}^{n}(x_{im} - \bar{x}_m)^2}\right)\left(\frac{1}{1 - R_j^2}\right), \tag{2.4}$$

där första termen är variansen av $\hat{\beta}_j$ i modellen $y_i = \alpha + \beta_j x_{ij} + \epsilon_i$ och $\frac{1}{1-R_j^2}$ är VIF-faktorn (Sundberg, 2014, s. 73). R_j^2 är förklaringsgraden som anger hur mycket av variationen i x_j som förklaras av de andra förklaringsvariablerna. I det fall då x_j är ortogonal mot de övriga variablerna så är $R_j^2 = 0$, vilket i sin tur medför att VIF $= 1$. Vi ser på formel (2.4) att inflationsfaktorn ökar ju större andel av variationen i x_j som kan förklaras av de andra variablerna. VIF anger alltså hur mycket variansen i regressionskoefficientskattaren $\hat{\beta}_j$ ökar då det råder korrelationer mellan förklarande variabler. Ett högt VIF värde indikerar på multikollinjäritet, där värdet 5 är satt till den kritiska gränsen (Sundberg, 2014, s. 74).

2.2.5 Akaikes informationskriterium - AIC

Akaikes informationskriterium, AIC, används vid modelljämförelser i samma datamängd. Det är ett mått på hur väl en modell passar data i jämförelse till en eller flera alternativa modeller och ges av

$$AIC = -2l(\hat{\theta}_{ML}) + 2m,$$

där $l(\hat{\theta}_{ML})$ är den maximerade log likelihood och m är antalet parameterar i modellen, inklusive interceptet. Ju lägre AIC, desto bättre modell. Dock säger inte AIC något om hur bra själva modellen är, utan används mest till att jämföra olika modeller (Held & Sabanés Bové, 2014, s. 224).

2.2.6 Residualer

Residualer är differensen mellan de observerade värdena y_i och regressionslinjen. För att kontrollera om en multipel linjär regressionsmodell anpassas väl med data kan vi undersöka residualerna genom visualisation. Två av förutsättningarna för linjär regression (se avsnitt 2.1) är att residualerna måste ha konstant varians och vara normalfördelade med väntevärde 0. Genom att plotta residualerna mot de skattade värdena $\hat{y}_i$ kan man se om variansen är konstant eller inte. Om variansen inte är konstant så följer residualerna ett systematiskt mönster, detta kallas även för heteroskedasticitet.

Då variansen är konstant är residualerna jämnt utspridda kring 0 och indikerar på att residualerna inte påverkas då värdet på responsvariabeln ökas (Andersson & Tyrcha, 2015, s. 63). För att undersöka om residualerna är normalfördelade kan man plotta en normalfördelningsplot. Om punkterna approximativt följer en rät linje kan residualerna antas vara normalfördelade (Andersson & Tyrcha, 2015, s. 63-64). I det fall då antaganden om residualerna inte uppfylls kan man utföra variabeltransformationer.

2.3 Transformering av variabler

I de fall då modellen inte är linjär eller inte uppfyller något av förutsättningarna för linjär regression kan det vara lämpligt att transformera variabler. En vanligt förekommande transformation är att logaritmera responsvariabeln eller de förklarande variablerna, eller både responsvariabeln och förklarande variabler. I följande delavsnitt presenteras de transformationer som vi kommer att tillämpa.

2.3.1 Dummyvariabel

Kategoriska variabler är variabler som har två eller flera kategorier. Ett enkelt exempel på en kategorisk variabel är kön, som har två kategorier; kvinna och man. Om de inte har värden som går att rangordnas så kan de ändå användas som regressor i en regressionsmodell om kategorierna omvandlas till dummyvariabler. Dessa variabler antar antingen 1 eller 0. Variabeln har värdet 1 då den innehar egenskapen man är intresserad av, annars är värdet 0. Om det finns n kategorier så skapas $n-1$ dummyvariabler, där den n:te kategorin används som referenskategori. Om vi till exempel har tre olika egenskaper i variabeln `Material`; glas, plast och metall och låter variabeln `Glas` anta värdet 1 då materialet är av glas, 0 annars och variabeln `Plast` anta värdet 1 om materialet är av plast, 0 annars, så är materialet av metall endast då både glas och plast antar värdet 0. Dummyvariabler kan även användas till att dela upp grupper. I vår datamängd finns många olika tillverkare. Vi kan då dela upp de i två olika grupper; populära och mindre populära tillverkare, där variabeln har värdet 1 då tillverkaren är populär och 0 då tillverkaren är mindre populär (Sundberg, 2014, s. 82-83).

2.3.2 Logtransformering

I det fall då sambandet mellan de förklarande variablerna och responsvariabeln inte är linjärt kan man bland annat logtransformera responsvariabeln för att hantera det (Sundberg, 2014, s. 83-84). Vi får då följande

$$\log(y_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \epsilon_i. \tag{2.5}$$

Man kan även logtransformera en eller flera förklarande variabler (2.6) eller både förklarande variabler och responsvariabeln (2.7)

$$y_i = \alpha + \beta_1 log(x_{1i}) + \beta_2 log(x_{2i}) + \ldots + \beta_m log(x_{mi}) + \epsilon_i \tag{2.6}$$

$$log(y_i) = \alpha + \beta_1 log(x_{1i}) + \beta_2 log(x_{2i}) + \ldots + \beta_m log(x_{mi}) + \epsilon_i \tag{2.7}$$

I det fall då vi logtransformerar responsvariabeln och sedan vill transformera tillbaka till en icke-logaritmerade responsvariabel fås

$$y_i = e^{\alpha} \cdot e^{\beta_1 x_{1i}} \cdot e^{\beta_2 x_{2i}} \cdot \ldots \cdot e^{\beta_m x_{mi}} \cdot \epsilon'$$

Där $\epsilon' = e^{\epsilon}$ och är log-normalfördelad med väntevärde $e^{\sigma^2/2}$ och varians $(e^{\sigma^2} - 1)e^{\sigma^2}$.

I detta arbete kommer vi att använda (2.5) och en version av (2.7), där vi logaritmerar responsvariabeln och endast en förklarande variabel.

2.3.3 Saknade värden

Det är vanligt att vissa värden saknas i ett dataset. Det finns flera olika sätt att hantera dessa och i vårt fall kommer vi att ta bort de observationer med saknade värden. För att försäkra oss om att dessa observationer inte sticker ut från mängden kan vi plotta de i en scatterplot och se hur de beter sig jämfört med de resterande observationerna. I avsnitt (3.2) kommer vi att hantera saknade värden.

2.4 Prediktion

Då vi har en välanpassad modell kan vi använda den för prediktion, det vill säga för att förutsäga priser på mobiltelefoner. Vi använder då våra skattade β-värden och ersätter x-variablerna med variabler från en ny datamängd och får en ny responsvektor y^*.

2.4.1 Mean squared error of prediction - MSEP

MSEP är ett mått på prediktionsfelets storlek. Låt y_i^* vara de observerade värdena i ett nytt datamaterial och $\hat{y}_i^*$ vara prediktionerna. Då ges det genomsnittliga kvadratfelet i prediktionen av

$$MSEP = \frac{1}{n^*} \sum_{i=1}^{n^*} (y_i^* - \hat{y}_i^*)^2,$$

där n^* är antalet observationer i det nya datamaterialet (Sundberg, 2014 s. 70).

Ett lägre MSEP-värde är att föredra, då det innebär att avståndet mellan det predikterade- och det verkliga värdet är mindre. MSEP kan dock vara missvisande om framtida observationer inte följer samma modell som data vi använt till att ta fram prediktionsmodellerna.

Ett ekvivalent mått är RMSEP (Root Mean Squared Error of Prediction), som är kvadratroten ur MSEP.

3 Datamaterial

Datamaterialet har hämtats från prisjakt [10] och från ett antal hemsidor med information om mobilspecifikationer [5] [4] [6] [2]. Prisjakt är en hemsida där man kan jämföra priser på en given produkt mellan olika butiker. 200 mobiltelefoner med operativsystemet Android har slumpats fram ur de 1013 mobiltelefoner med Android som fanns på prisjakts hemsida. För att slumpa fram dessa 200 mobiltelefonerna har vi använt oss av programmeringsspråket `R`. Samtliga mobiltelefoner med operativsystemet Android på prisjakt lades i en dataframe. En kolumn med observationernas index lades till. Sedan lät vi `R` slumpa fram 200 nummer ur 1013 och därefter valdes de observationerna med index som motsvarar slumptalen som `R` slumpade fram. Fem variabler som finns med i vår datamängd var tillgängliga på prisjakt och kunde enkelt hämtas med hjälp av `R`. Information om de resterande variablerna har sedan hämtats manuellt från olika hemsidor med information om mobilspecifikationer. Av de variablerna som fanns valdes de variabler som tros ha inverkan på priset och som varierar mellan olika mobiltelefoner. Till exempel väljer vi bort variablerna Wi-fi (om mobiltelefonen stödjer Wi-fi eller inte) och högtalare (om mobiltelefonen har högtalare eller inte), då de flesta smarta mobiltelefoner idag stödjer Wi-fi och har högtalare. Det kan vara så att observationerna inte är oberoende i priset då till exempel om en telefon från Samsung kostar en viss summa, så kanske Sony prissätter sina mobiltelefoner utifrån det. Men oavsett om observationerna är oberoende eller ej så kommer vi att anta att de är oberoende.

3.1 Beskrivning av variabler

Responsvariabeln och de förklarande variablerna vi valt att ha med sammanställs i Tabell 1 nedan.

Variabel	Beskrivning
Pris	Priset som mobiltelefonen säljs för
Tillverkare	Mobilens tillverkare
Lansering	Vilket år mobiltelefonen lanserades
ProcessorTyp	Antalet kärnor processorn har
Skärmstorlek	Skärmens diagonallängd, mätt i tum
Skärmupplösning	Skärmens upplösning, mätt i pixlar
Bakkamera	Bakkamerans upplösning, mätt i megapixlar
Framkamera	Framkamerans upplösning, mätt i megapixlar
RAM	Primärminnets storlek, mätt i gigabyte
InterntMinne	Mobiltelefonens interna minne, mätt i gigabyte
Minneskort	Om mobiltelefonen har stöd för minneskort
Vattentålig	Om mobiltelefonen är vattentålig
NFC	Om mobiltelefonen har stöd för NFC
Fingeravtryck	Om mobiltelefonen har fingeravtrycksläsare
Längd	Mobiltelefonens längd, mätt i mm
Bredd	Mobiltelefonens bredd, mätt i mm
Djup	Mobiltelefonens tjocklek, mätt i mm
Vikt	Hur mycket mobiltelefonen väger, mätt i gram
Batterikapacitet	Batteriets kapacitet, mätt i mAh
UtbytbartBatteri	Om man kan byta ut batteriet
Material	Vilket material mobiltelefonen är gjord av (aluminium, glas, metall, plast, polykarbonat eller läder).

Tabell 1: Beskrivning av förklarande variabler.

3.2 Transformationer och behandling av variabler

Mobiltelefoner av samma modell lanseras oftast med olika stora interna minnen, så de som behöver mer minne har möjligheten att köpa en mobiltelefon som har mycket plats i minnet. Varianter med mer minne kostar vanligtvis mer, även om resterande specifikationer är exakt lika. Detta medför att vi möjligtvis får dubbla observationer, det vill säga 2 eller fler mobiltelefoner av samma tillverkare i samma modell finns med i datamängden, där det enda som skiljer de åt är storleken på det interna minnet. En översikt av datamaterialet visade att det fanns 5 dubbla observationer, vi väljer att ta med de med lägst interna minne. Vi ser även att det finns 16 observationer som saknar värden i flera variabler. För att se om priserna på dessa observationer sticker ut ur mängden så plottar vi alla observationer mot responsvariabeln och markerar de observationer som saknar värden. Det visar sig att de inte beter sig annorlunda jämfört med de andra (se Figur 9 i appendix), så vi väljer att utesluta dessa observationer ur datamängden.

Eftersom att vi ska försöka hitta en modell som kan prediktera priset väljer vi att "lägga undan" de mobiltelefoner som lanserats år $2017 - 2018$ och tar fram de igen när vi ska testa modellen.

Ibland kan vidare analyser bli effektivare av att transformera variabler. Då en variabel antar numeriska värden kommer modellen att skatta endast en parameter. Om variabeln är kategorisk skattar modellen antalet parametrar som motsvarar antalet kategorinivåer. För att spara på data så är, i vårt fall, numeriska variabler att föredra, då de endast skattar en parameter och är effektivast. Vi ser att det finns 2 kategoriska variabler i datamängden som går att rankas, `Skärmupplösning` och `ProcessorTyp`, därmed kan vi göra en numerisk omskrivning av värdena. Eftersom att `single core, dual core` osv står för en-kärnig processor, två-kärnig processor osv så tilldelar vi de motsvarande numeriska värden. Skärmupplösning anges i antalet horisontella- och vertikala punkter. Vi tilldelar `Skärmupplösning` numeriska värden som motsvarar totala antalet pixlar och dividerar med 1000 för att inte få allt för höga värden. Variablerna har då följande värden:

Skärmupplösning	Numeriskt värde
$240x320$	76.8
$240x430$	103.2
$320x480$	153.6
$\vdots$	$\vdots$
$3120x4160$	12979.2

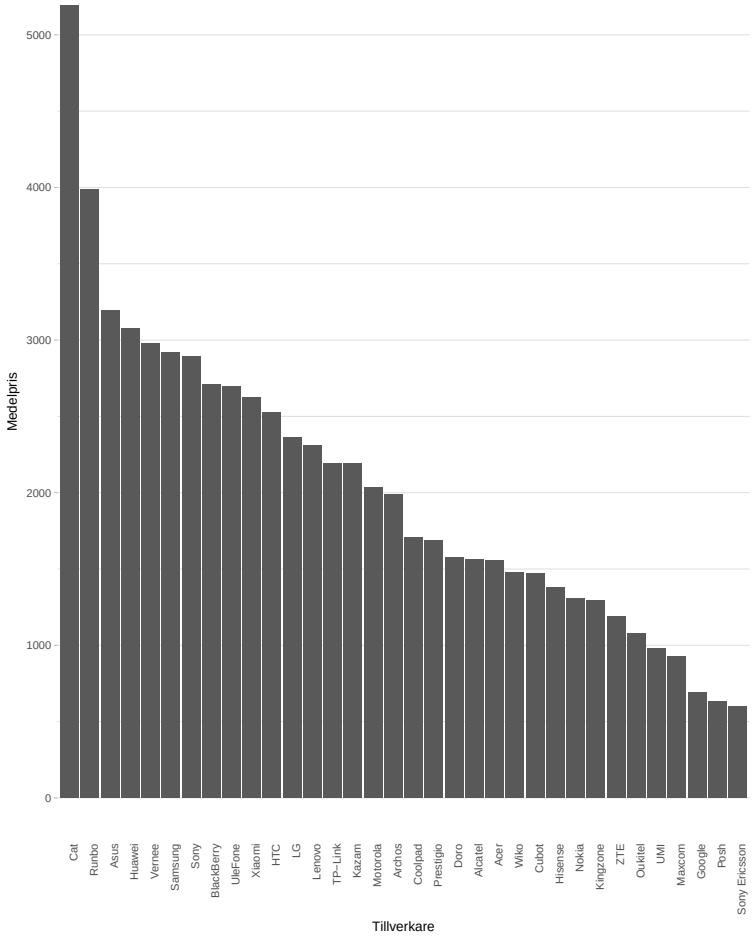
Tabell 2: Transformation av variabeln `Skärmupplösning`. Det numeriska värdet är totala antalet pixlar dividerat med 1000.

ProcessorTyp	Numeriskt värde
single core	1
dual core	2
quad core	4
hexa core	6
octa core	8
deca core	10

Tabell 3: Transformation av variabeln `ProcessorTyp`. Det numeriska värdet motsvarar antalet kärnor processorn har.

De två variablerna `Tillverkare` och `Material` är också kategoriska, men går inte att rankas. Vi väljer istället att transformera dem till dummyvariabler på följande vis:

Tillverkare: Antar värdet 1 om tillverkaren är Samsung, Huawei eller Sony. Antar värdet 0 annars. Vi tolkar det som att Samsung, Huawei och Sony är de populära mobiltelefonerna och de resterande är mindre populära. Vi väljer just dessa tillverkare då operatörerna Tele2, Telia, Telenor och Comviq endast säljer dessa. I Figur 1 ser vi skillnaderna på medelpriset hos olika tillverkare. Den kan dock vara missvisande då vi inte har särskilt stor datamängd att utgå ifrån.



Figur 1: Medelpriset hos olika tillverkare, baserat på de 200 mobiltelefonerna som slumpats ut.

Material: Material innehåller 6 olika kategorier; aluminium, glas, metall, plast, polykarbonat och läder. Eftersom polykarbonat är en typ av plast [9]

sätter vi ihop plast och polykarbonat som en gemensam kategori och kallar den för `plast`. Vi sätter även ihop aluminium och metall till en gemensam kategori och kallar den för `metall`, då aluminium är en typ av metall [8]. Vi har då bara 4 kategorier i variabeln `Material`; metall, plast, glas och läder. Vi låter läder vara vår referensnivå, då läder förekommer minst bland observationerna. Vi skapar följande dummyvariabler:

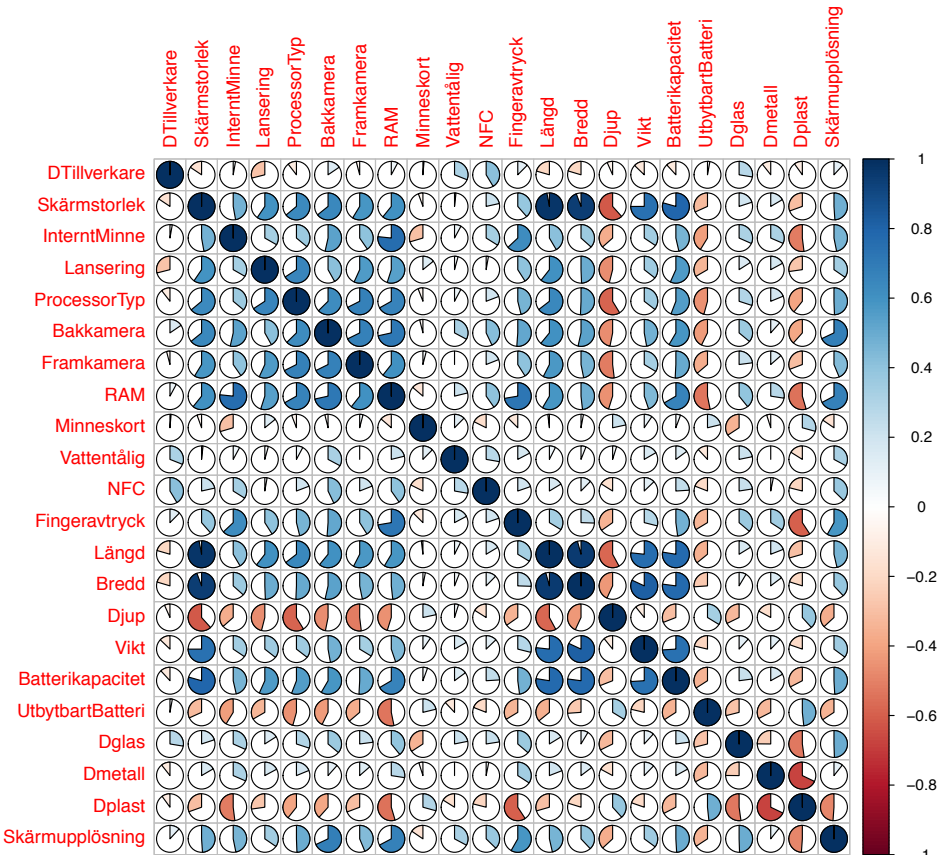
Metall: 1 om materialet är av metall, 0 annars.

Plast: 1 om materialet är av plast, 0 annars.

Glas: 1 om materialet är av glas, 0 annars.

Om alla kategorierna är 0, så är materialet av läder.

Vidare undersöker vi multikollinjäritet med VIF-värdena. Vi ser i Tabell 13 i appendix att `RAM`, `Skärmstorlek`, `Längd`, `Bredd`, `Djup`, `Vikt` och samtliga material har höga värden. För att se hur våra variabler är korrelerade med varandra gör vi en korrelationsmatris med hjälp av `R`.



Figur 2: Korrelationsmatris mellan de numeriska och dummy förklarande variablerna.

Vi ser på korrelationsmatrisen i Figur 2 att det är hög korrelation mellan `Skärmstorlek` och `Längd`, `Skärmstorlek` och `Bredd` samt `Längd` och `Bredd`. Vi transformerar våra variabler i försök om att sänka VIF värdena och åtgärda korrelationerna. Vi ersätter variablerna `Längd`, `Bredd` och `Djup` med den nya variabeln `Volym` = (Längd · Bredd · Djup)/1000 med enheten cm^3 och sedan ersätter vi variablerna `Skärmupplösning` och `Skärmstorlek` med den nya variabeln `Skärm = Skärmupplösning / Skärmstorlek`, alltså skärmupplösning per skärmstorlek med enheten 1000 pixlar/tum. Vi tittar närmare på korrelationsvärden mellan variabeln `RAM` och de resterande variablerna i Tabell 4.

	RAM
DTillverkare	0.08
InterntMinne	**0.77**
Lansering	0.55
ProcessorTyp	0.67
Bakkamera	**0.72**
Framkamera	0.61
Minneskort	−0.14
Vattentålig	0.21
NFC	0.40
Fingeravtryck	**0.72**
Volym	0.14
Vikt	0.45
Batterikapacitet	0.67
UtbytbartBatteri	−0.53
Skärm	0.65
Dglas	0.39
Dmetall	0.27
Dplast	−0.55

Tabell 4: Korrelation mellan `RAM` och resterande variabler.

Vi ser att `RAM` har en korrelation över 0.7 med `InterntMinne`, `Bakkamera` och `Fingeravtryck`. `RAM` är alltså högt korrelerad med dessa och kan skapa missvisande resultat i kommande analysen, så vi väljer att utesluta `RAM` tills vidare.

Efter denna transformation visar det sig att endast `Dplast`, `Dglas` och `Dmetall` har höga VIF-värden som överstiger den kritiska gränsen (se Tabell 14 i appendix). Detta var väntat då vi skapat dessa dummyvariabler utifrån en kategorisk variabel. Vi kan dock bortse från dessa höga VIF-värden, då vi vet att om materialet är av plast så kan det omöjligen vara av till exempel glas.

3.3 Slutgiltiga variabler

Följande variabler är de vi kommer att använda oss av i den kommande analysen.

Variabel	Typ	Min.värde	Max.värde
Pris (respons var.)	Numerisk	389(kronor)	9399(kronor)
DTillverkare	Dummy	0	1
InterntMinne	Numerisk	0.128(gigabyte)	128(gigabyte)
Lansering	Numerisk	2010	2016
ProcessorTyp	Numerisk	1	6
Bakkamera	Numerisk	2(megapixlar)	23(megapixlar)
Framkamera	Numerisk	0(megapixlar)	16(megapixlar)
Minneskort	Dummy	0	1
Vattentålig	Dummy	0	1
NFC	Dummy	0	1
Fingeravtryck	Dummy	0	1
Volym	Numerisk	4859.797(cm^3)	22680.072(cm^3)
Vikt	Numerisk	52.7(gram)	250(gram)
Batterikapacitet	Numerisk	650(milli Ampere timmar)	5100(milli Ampere timmar)
UtbytbartBatteri	Dummy	0	1
Skärm	Numerisk	25.6(1000pixlar/tum)	1508.07(100pixlar/tum)
Dglas	Dummy	0	1
Dplast	Dummy	0	1
Dmetall	Dummy	0	1

Tabell 5: Slutgiltiga variabler efter variabeltransformationer.

4 Statistisk modellering

4.1 Enkel linjär regression

Vi börjar med att utföra enkel linjär regression av varje variabel enskilt för att se hur väl de förklarar responsvariabeln ensamt.

Variabel	Estimat	P-värde	R^2
DTillverkare	808	$5.38 \cdot 10^{-3}$	0.0585
InterntMinne	55.49	$< 2.2 \cdot 10^{-16}$	0.4449
Lansering	387.3	$3.15 \cdot 10^{-5}$	0.126
ProcessorTyp	575.2	$5.53 \cdot 10^{-8}$	0.205
Bakkamera	201.7	$1.54 \cdot 10^{-15}$	0.39
Framkamera	214.4	$8.80 \cdot 10^{-7}$	0.17
Minneskort	-1360.1	$4.29 \cdot 10^{-3}$	0.062
Vattentålig	1725.9	$5.55 \cdot 10^{-5}$	0.1188
NFC	1416	$2.61 \cdot 10^{-7}$	0.186
Fingeravtryck	2164.9	$4.04 \cdot 10^{-14}$	0.359
Vikt	23.772	$1.31 \cdot 10^{-6}$	0.1666
Volym	0.1176	$9.87 \cdot 10^{-2}$	0.021
Batterikapacitet	0.9970	$5.31 \cdot 10^{-10}$	0.259
UtbytbartBatteri	-1375.5	$6.40 \cdot 10^{-7}$	0.175
Dplast	-1615.8	$4.32 \cdot 10^{-9}$	0.24
Dglas	2313	$2.45 \cdot 10^{-10}$	0.2678
Dmetall	352.3	$2.943 \cdot 10^{-1}$	0.0085
Skärm	4.8912	$< 2.2 \cdot 10^{-16}$	0.47

Tabell 6: Estimater, p-värden och R^2 för varje enskild förklarande variabel i enkel linjär regression.

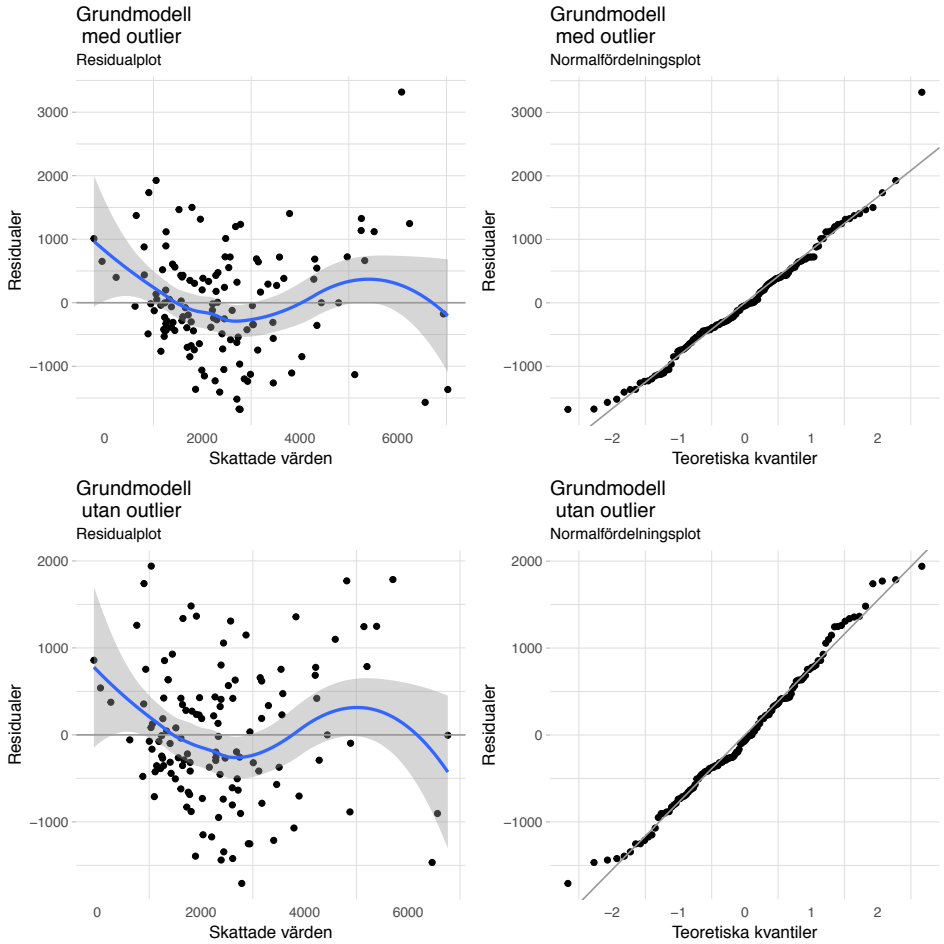
Utifrån Tabell 6 ser vi att alla variabler förutom `Dmetall` och `Volym` har en signifikant inverkan, på nivån 0.05, på responsvariabeln och `Skärm` är variabeln som förklarar priset bäst med $R^2 = 0.47$.

4.2 Multipel linjär regression

Grundmodellen som vi kommer att titta på först innehåller samtliga variabler. För att se om denna modell är lämplig för regression, det vill säga uppfyller villkoren i avsnitt 2.1, undersöker vi modellens residualer. Vi vet redan från tidigare beräkningar att vi inte har problem med multikollinjäritet, se avsnitt 3.2.

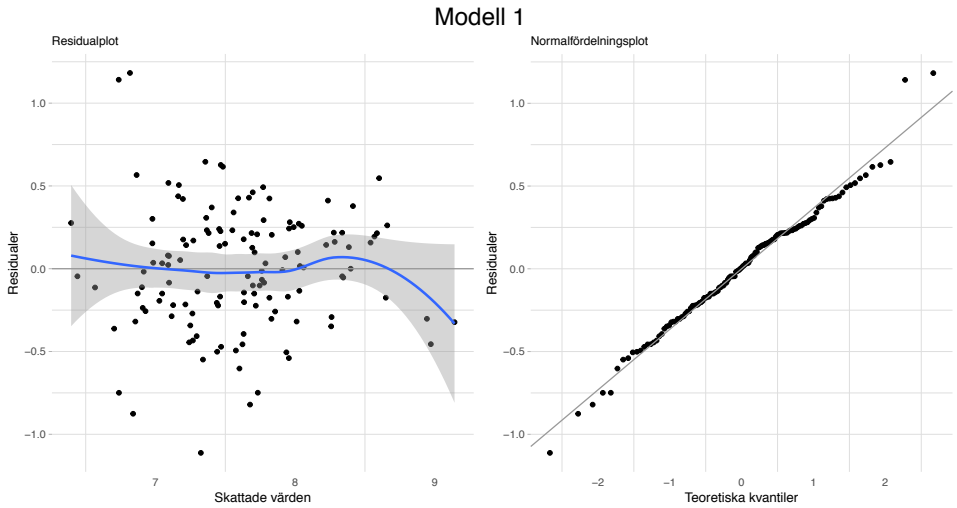
Vi ser på första raden i Figur 3 att fördelningen för residualerna tenderar till normalfördeling, men residualerna ser inte ut att ha konstant varians då punkterna inte är jämt fördelade kring noll i residualplotten. För att försöka åtgärda detta gör vi transformationer av variablerna. I ett första försök undersöker vi om det finns några potentiella outliers som påverkar plottens utseende. På första raden i Figur 3 ser vi att det finns en potentiell outlier. Vi tar bort denna observation och undersöker residualerna. På andra raden i samma figur ser vi att det inte blir någon skillnad förutom att en

observation är borta, så det är fortfarande heteroskedastiskt. En närmare titt på denna observation visar att den har högsta priset, men skiljer sig inte så mycket från resterande och kan då inte antas vara felaktig. R^2 påverkas inte av elimineringen. Vi väljer att behålla observationen och fortsätter med transformering.



Figur 3: Residualplot samt normalfördelningsplot för grundmodellen med och utan outlier.

Vidare undersöker vi residualerna då responsvariabeln logaritmeras. Denna modell kallar vi för modell 1 och innehåller samtliga variabler med logaritmerad responsvariabel. Residualerna illustreras i Figur 4 nedan.



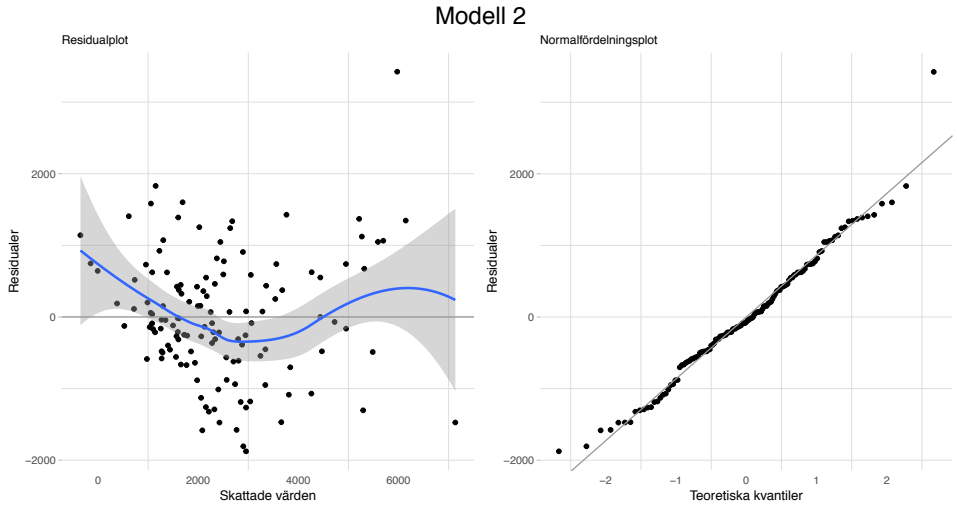
Figur 4: Residualplot samt normalfördelningsplot för modell 1.

I Figur 4 syns det att residualerna är jämt fördelade kring noll, alltså kan vi anta att de har konstant varians. Tittar vi på normalfördelningsplotten ser det ut som att residualerna tenderar till normalfördelning. Det går att se i residualplotten att det finns tre potentiella outliers. Borttagning av dem visade inga markanta skillnader och de verkar inte vara felaktiga. R^2 har inte påverkats. Vi väljer att ha kvar dem.

Eftersom att förklaringsvariabeln `Skärm` hade högst förklaringsgrad när vi utförde enkel linjär regression i avsnitt 4.1, så förklarar den variationen i responsvariabeln mest. `Skärm` är alltså den förklarande variabeln som kan påverka plottens utseende mest vid transformation. Därmed undersöker vi två nya modeller med dessa transformationer

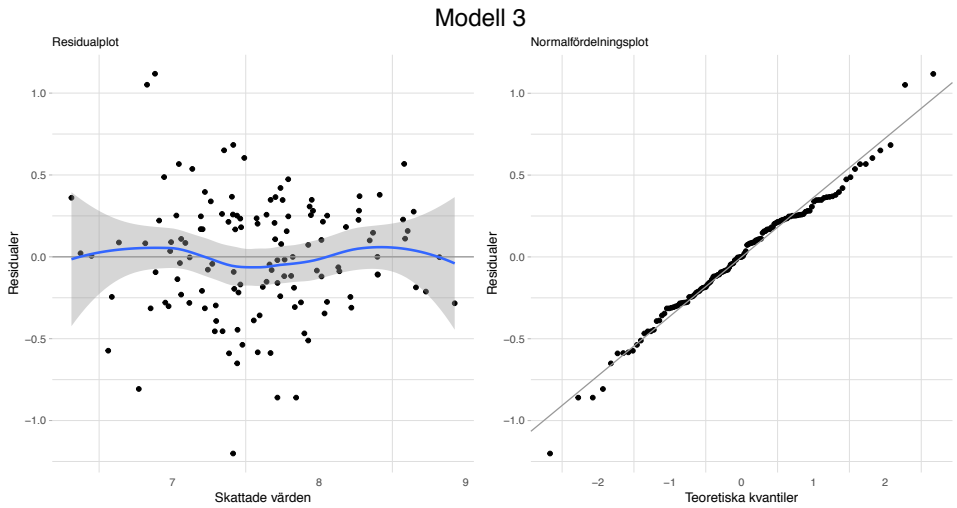
- logaritmerad `Skärm`
- logaritmerad `Pris` och logaritmerad `Skärm`

Vi börjar med att undersöka modell 2, som innehåller samtliga variabler, men med logaritmerad `Skärm`.



Figur 5: Residualplot samt normalfördelningsplot för modell 2.

Utifrån Figur 5 kan vi konstatera att residualerna är heteroskedastiska då punkterna inte är jämt fördelade kring noll och följer ett tydligt mönster. Borttagning av de potentiella outliers gav ingen förbättring. De verkar inte vara felaktiga och R^2 påverkas inte av eliminationen, så vi behåller dem för tillfället. Normalfördelningsplotten visar att residualerna tenderar till normalfördelning. Slutligen tittar vi på residualerna för modell 3, som innehåller samtliga variabler med logaritmerad responsvariabel och logaritmerad `Skärm`.



Figur 6: Residualplot samt normalfördelningsplot för modell 3.

Vi kan se i Figur 6 att residualerna i modell 3 ser ut att ha konstant varians och tenderar till normalfördelning. Även i denna modell syns tre potentiella outliers, men vi behåller dem av samma anledning som de första modellerna.

Vi kan alltså konstatera att modell 1 och modell 3 uppfyller förutsättningarna för linjär regression och vi kommer därför jämföra dessa två modeller i följande avsnitt.

4.2.1 Stegvis variabelselektion

För att se om det finns förklarande variabler med koefficienter som inte är signifikanta skild från noll använder vi oss av samtliga variabelselektionsmetoder som togs upp i avsnitt 2.2.2. Det visar sig att alla tre metoderna ger samma resultat på respektive modell. De reducerade modellerna med endast signifikanta variabler är som följande

Reducerad modell 1

$$\begin{aligned} logPris_i = \alpha + \beta_1 Skarm + \beta_2 InterntMinne + \beta_3 Vikt + \\ \beta_4 Vattentalig + \beta_5 DTillverkare + \beta_6 Lansering \end{aligned} \tag{4.1}$$

Reducerad modell 3

$$\begin{aligned} logPris_i = \alpha + \beta_1 logSkarm + \beta_2 InterntMinne + \beta_3 Vikt + \\ \beta_4 Vattentalig + \beta_5 DTillverkare + \beta_6 Lansering \end{aligned} \tag{4.2}$$

Koefficienterna visas i Tabell 9 och 11.

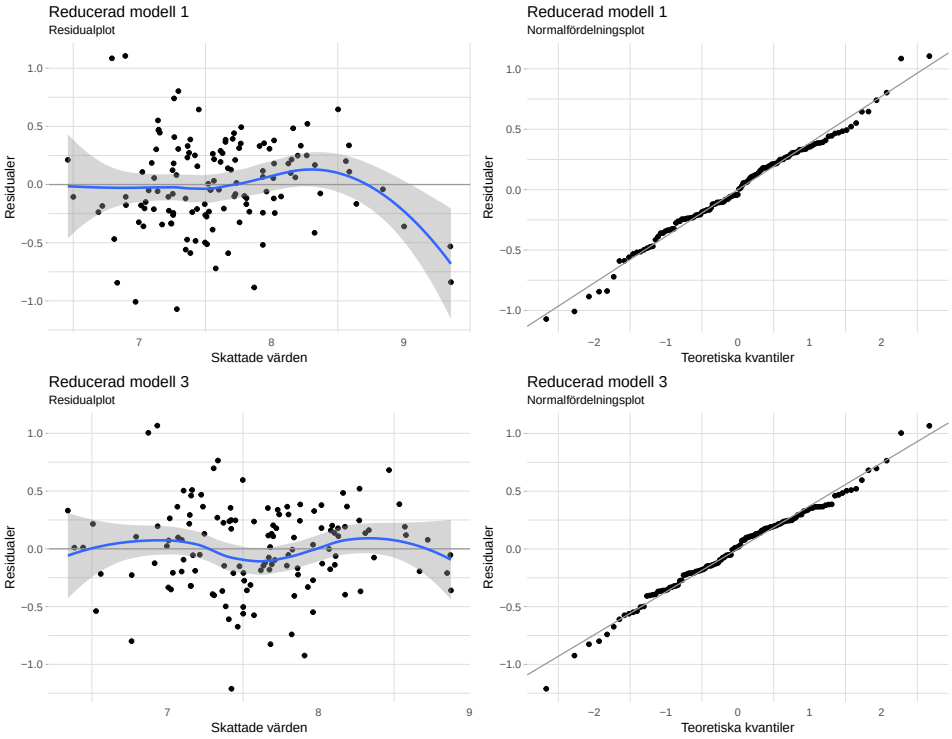
4.3 Modelljämförelse

Innan vi kan dra några slutsatser om dessa modeller undersöker vi och jämför modellernas R^2_{adj}, AIC och residualernas standardfel. I Tabell 7 nedan ser vi en jämförelse mellan dessa för modell 1 och modell 3 efter variabelselektion.

	R^2_{adj}	AIC	Residual standardfel
Reducerad modell 1	0.6366	137.2	0.3950
Reducerad modell 3	0.6621	127.6	0.3808

Tabell 7: Jämförelse mellan R^2_{adj}, AIC och residualernas standardfel i regressionsmodell 1 och 3 efter variabelselektion.

I Tabell 7 ovan ser vi att reducerad modell 3 har ett R^2_{adj}-värde högre än reducerad modell 3 och AIC och residualernas standardfel är något lägre i reducerad modell 3 än i reducerad modell 1. Vi gör även en jämförelse mellan residualerna på reducerad modell 1 och reducerad modell 3.



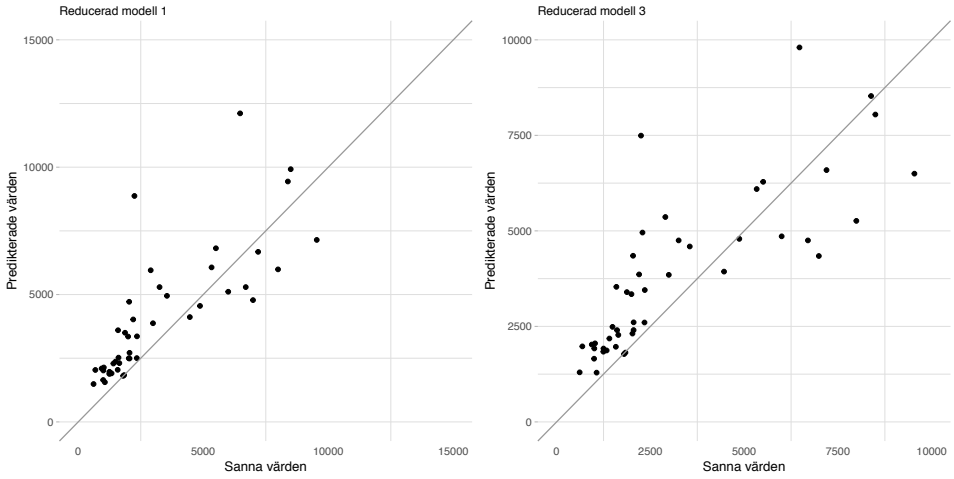
Figur 7: Residualplot samt normalfördelningsplot för modell 1 och modell 3 efter variabelselektion.

Vi ser i Figur 7 att residualerna ser ut att ha konstant varians och vara normalfördelade i båda modellerna. Utifrån dessa jämförelser kan vi då konstatera att reducerad modell 3 passar vår data bäst.

4.4 Prediktion

Som tidigare nämnt har vi använt oss av datamängden med lanseringsdatum mellan 2010 – 2016 i regressionsanalysen och lagt undan datamängden med lanseringsdatum 2017 – 2018, i syfte om att testa modellens prediktionsförmåga. Vi jämför prediktionsförmågan på reducerad modell 1 och reducerad modell 3 genom att tillämpa modellerna på datamängden vi la undan och undersöker sedan MSEP- och RMSEP-värden.

Vi börjar med att illustrera sanna värden mot predikterade värden. Om predikterade värden skulle vara approximativt lika med sanna värden så skulle de följa en rät linje genom origo.



Figur 8: Sanna värden mot predikterade värden för reducerad modell 1 och reducerad modell 3.

Vi ser i Figur 8 att predikterade värden inte följer längs den räta linjen helt. Ju högre pris, desto mer utspridd är punkterna, men det ser ändå ut som att de approximativt följer linjens riktning. Vi använder oss av måttet MSEP och RMSEP för att få ett värde på hur stora prediktionsfelen är. Värdena för båda modellerna presenteras i Tabell 8.

	$MSEP$	RMSEP
Reducerad modell 1	11657983	3414.4
Reducerad modell 3	2698144	1642.6

Tabell 8: MSEP och RMSEP för reducerad modell 1 och reducerad modell 3.

Eftersom att lägre MSEP- och RMSEP-värde är önskevärd kan vi utifrån Tabell 8 konstatera att reducerad modell 3 är den modell som predikterar framtida priser bäst av dessa två modeller.

5 Resultat

Av de fyra modellerna som vi jämförde mellan, har vi kommit fram till att reducerad modell 1 och reducerad modell 3 uppfyllde förutsättningarna för linjär regression. Det visade sig att reducerad modell 3 passade data bäst och är modellen som är bäst lämpad för prediktion av de två modellerna. Då vi logaritmerade responsvariabeln på båda modellerna så behöver vi omvandla högra och vänstra leden till potenser av e för att få en icke-logaritmerad

responsvariabel. Vi får då multiplikativa faktorer, med koefficienterna som presenteras i Tabell 9 och Tabell 11;

Reducerad modell 1

$$Pris_i = e^{\hat{\alpha}} \cdot e^{\hat{\beta}_1 Skarm} \cdot e^{\hat{\beta}_2 InterntMinne} \cdot e^{\hat{\beta}_3 Vikt} \cdot e^{\hat{\beta}_4 Vattentalig} \cdot e^{\hat{\beta}_5 DTillverkare} \cdot e^{\hat{\beta}_6 Lansering} \cdot \epsilon' \tag{5.1}$$

Reducerad modell 3

$$Pris_i = e^{\hat{\alpha}} \cdot e^{\hat{\beta}_1 logSkarm} \cdot e^{\hat{\beta}_2 InterntMinne} \cdot e^{\hat{\beta}_3 Vikt} \cdot e^{\hat{\beta}_4 Vattentalig} \cdot e^{\hat{\beta}_5 DTillverkare} \cdot e^{\hat{\beta}_6 Lansering} \cdot \epsilon' \tag{5.2}$$

Här är $\epsilon' = e^{\hat{\epsilon}}$ log-normalfördelad med väntevärde $e^{\sigma^2/2}$ och varians $(e^{\sigma^2} - 1)e^{\sigma^2}$. Reducerad modell 1 har en justerad förklaringsgrad på 0.6366 och MSEP-värde på 11657983. Den justerade förklaringsgraden för reducerad modell 3 är 0.6621 och MSEP-värdet är 2698144. Eftersom att faktorerna nu är multiplikativa så ökar responsen multiplikativt med e^{β_i} då x_i förändras med en enhet. Man kan även istället för att se hur mycket responsen ökar multiplikativt, se hur mycket den ökar procentuellt genom att dra bort 1 från $e^{\beta_i x_i}$ och multiplicera med 100. Vi tolkar skattningarna för respektive modell i följande delavsnitt.

5.1 Reducerad modell 1

Estimater, p-värden och standardavvikelser för reducerad modell 1 sammanställs i Tabell 9 nedan.

Koefficient	Estimat	Standardavvikelse	P-värde
Intercept	$-1.899 \cdot 10^2$	$5.468 \cdot 10^1$	$7.12 \cdot 10^{-4}$
Skärm	$9.327 \cdot 10^{-4}$	$1.840 \cdot 10^{-4}$	$1.42 \cdot 10^{-6}$
InterntMinne	$9.817 \cdot 10^{-3}$	$2.062 \cdot 10^{-3}$	$5.27 \cdot 10^{-6}$
Vikt	$4.574 \cdot 10^{-3}$	$1.389 \cdot 10^{-3}$	$1.29 \cdot 10^{-3}$
Vattentålig	$2.357 \cdot 10^{-1}$	$1.190 \cdot 10^{-1}$	$4.99 \cdot 10^{-2}$
DTillverkare	$2.006 \cdot 10^{-1}$	$7.977 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$
Lansering	$9.728 \cdot 10^{-2}$	$2.715 \cdot 10^{-2}$	$4.87 \cdot 10^{-4}$

Tabell 9: Koefficienter, estimater, standardavvikelser och p-värde för reducerad modell 1.

Interceptet $\hat{\alpha}$ har ett värde på $e^{-0.01899} = 0.98$ och är punkten där regressionlinjen korsar y-axeln. Vi kommer inte att fördjupa oss i det, då det inte är intressant i detta fall. β-skattningarna för de numeriska variablerna `Skärm`,

`InterntMinne`, `Vikt` och `Lansering` tolkar vi som att en förändring i variabel x_i ger priset en multiplikativ ökning med e^{β_i}. Till exempel ger en enhets förändring i `InterntMinne` en prisökning med faktorn $e^{0.009817} = 1.0099$, det vill säga priset ökas med 0.99%. β-skattningen för dummyvariabeln `Vattentålig` tolkas som att en förändring i variabeln ger priset en multiplikativ ökning med $e^{0.2357} = 1.266$ (26.6%). På liknande sätt tolkar vi dummyvariabeln `DTillverkare`. Det visar sig att vattentåligheten ger mest effekt på priset. Med en enhets förändring i `Vattentålig` ökar priset multiplikativt med $e^{0.2357} = 1.266$ (26.6%). Variabeln som ger minst förändring på priset är skärmupplösning per skärmstorlek. Priset ökar med faktorn $e^{0.0009327} = 1.00093$ (0.09%) för varje 1000 pixlar/tum. Samtliga variabler i modellen ger en signifikant positiv effekt på `Pris`, se Tabell 9.

De variabler som visade sig vara icke-signifikanta i variabelselektionen presenteras i Tabell 10, med respektive p-värde innan variabelselektionen.

Variabel	P-värde
ProcessorTyp	0.42494
Bakkamera	0.85739
Framkamera	0.38426
Minneskort	0.17962
NFC	0.76634
Fingeravtryck	0.68561
Volym	0.57314
Batterikapacitet	0.41821
UtbytbartBatteri	0.26600
Dplast	0.25060
Dglas	0.29726
Dmetall	0.25510

Tabell 10: P-värde för variabler i modell 1 innan variabelselektion, som efter variabelselektion visade sig vara icke-signifikanta.

5.2 Reducerad modell 3

Estimater, p-värden och standardavvikelser för reducerad modell 3 sammanställs i Tabell 11 nedan.

Koefficient	Estimat	Standardavvikelse	P-värde
Intercept	$-1.350 \cdot 10^{2}$	$5.482 \cdot 10^{1}$	$1.51 \cdot 10^{-2}$
DTillverkare	$2.038 \cdot 10^{-1}$	$7.673 \cdot 10^{-2}$	$8.93 \cdot 10^{-3}$
InterntMinne	$8.233 \cdot 10^{-3}$	$2.052 \cdot 10^{-3}$	$1.03 \cdot 10^{-4}$
Lansering	$6.934 \cdot 10^{-2}$	$2.726 \cdot 10^{-2}$	$1.22 \cdot 10^{-2}$
Vattentålig	$2.711 \cdot 10^{-1}$	$1.122 \cdot 10^{-1}$	$1.72 \cdot 10^{-2}$
Vikt	$3.788 \cdot 10^{-3}$	$1.355 \cdot 10^{-3}$	$5.998 \cdot 10^{-3}$
logSkärm	$3.493 \cdot 10^{-1}$	$5.743 \cdot 10^{-2}$	$1.36 \cdot 10^{-8}$

Tabell 11: Koefficienter, estimater, standardavvikelser och p-värde för reducerad modell 3.

Interceptet $\hat{\alpha}$ har ett värde på $e^{-0.0135} = 0.99$ och är punkten där regressionlinjen korsar y-axeln. Vi kommer inte att fördjupa oss i det, då det inte är intressant i detta fall. β-skattningarna för de numeriska variablerna `InterntMinne`, `Lansering`, `Vikt` och `logSkärm` tolkar vi som en enhets förändring i variabel x_i ger priset en multiplikativ förändring med faktorn e^{β_i}. β-skattningen för dummyvariabeln `Vattentålig` tolkas som en förändring i variabeln ger priset en multiplikativ ökning med $e^{0.2711} = 1.311$ (31.1%). På liknande sätt tolkar vi dummyvariabeln `DTillverkare`. Det visar sig att även i reducerad modell 3 är `logSkärm` den variabeln som ger mest ökning på priset per enhetsförändring. Priset ökar med faktorn $e^{0.3493} = 1.42$ (42%) för varje enhetsändring i `logSkärm`. Mobiltelefonens vikt har minst effekt på priset. Priset ökar med faktorn $e^{0.003788} = 1.0038$ per gram, det vill säga 0.38%. Samtliga variabler i modellen ger en signifikant positiv effekt på priset, se Tabell 11.

De variabler som visade sig vara icke-signifikanta i variabelselektionen presenteras i Tabell 12, med respektive p-värde innan variabelselektionen.

Variabel	P-värde
ProcessorTyp	0.86428
Bakkamera	0.94220
Framkamera	0.53579
Minneskort	0.32104
NFC	0.89492
Fingeravtryck	0.96970
Volym	0.44791
Batterikapacitet	0.82417
UtbytbartBatteri	0.36795
Dplast	0.28170
Dglas	0.38805
Dmetall	0.26762

Tabell 12: P-värde för variabler i modell 3 innan variabelselektion, som efter variabelselektion visade sig vara icke-signifikanta.

Som det nämnts tidigare passar reducerad modell 3 data bäst och är även lämpligare i prediktionssyfte, då den har lägre MSEP- och RMSEP-värde än reducerad modell 1. Reducerad modell 3 är alltså vår slutmodell. Vid jämförelse av Tabell 10 och Tabell 12 ser vi att estimaterna är någorlunda lika för samtliga koefficienter förutom `Skärm`, men i reducerad modell 3 är `Skärm` logaritmerad.

6 Diskussion

Vi har i detta arbete undersökt vilka variabler som har ett statistiskt samband med priset på en smart mobiltelefon med operativsystemet Android. Reducerad modell 1 och reducerad modell 3 uppfyllde förutsättningarna för linjär regression och innehåller samma signifikanta variabler. Det som skiljer dem åt är att `Skärm` är logaritmerad i sistnämnda modellen och denna modell har även högre justerad förklaringsgrad och lägre AIC. Utifrån det kom vi fram till att reducerad modell 3 blev vår slutmodell. Den passade vår data bäst och var även modellen med lägst MSEP-värde, med andra ord är denna modell bättre lämpad för prediktion.

Slutmodellen förklarar bara 66% av den totala variationen i `Pris`, vilket betyder att ungefär en tredjedel av variationen inte förklaras. Tittar vi på grundmodellen, som innehåller alla 18 variabler, så förklarar den 70% av den totala variationen i `Pris`. Detta kan vara förklaringen till varför slutmodellen har låg justerad förklaringsgrad. Det visade sig att bland annat storleken på interna minnet och vilket år mobiltelefonen lanserades hade positiv effekt på priset. Detta är inte förvånansvärt, då man är villig att

betala mer för mer minne och nya mobiltelefoner är oftast dyrare än äldre, men blir billigare senare med åren. Även vikt hade positiv effekt. Man kan tänka sig att ju mer material och delar som används till att bygga upp mobiltelefonen, desto mer väger den och desto högre bli kostnaden. Tillverkare gav också postivit effekt, detta kanske inte är så förvånansvärt då populära mobiltelefoner säljer mer än mindre populära mobiltelefoner. Men man kan tycka att tillverkare inte borde ha någon effekt på priset, då en populär mobiltelefon som har samma specifikationer som en mindre populär mobiltelefon borde kosta lika mycket. Två variabler som vi trodde skulle ha en effekt på priset, men som visade sig att de inte var det är `Fingeravtryck` och `ProcessorTyp`. Det krävs mer avancerad teknik för att kunna låsa upp skärmen med sitt fingeravtryck och snabbare processorer är dyrare, därav borde det även kosta mer.

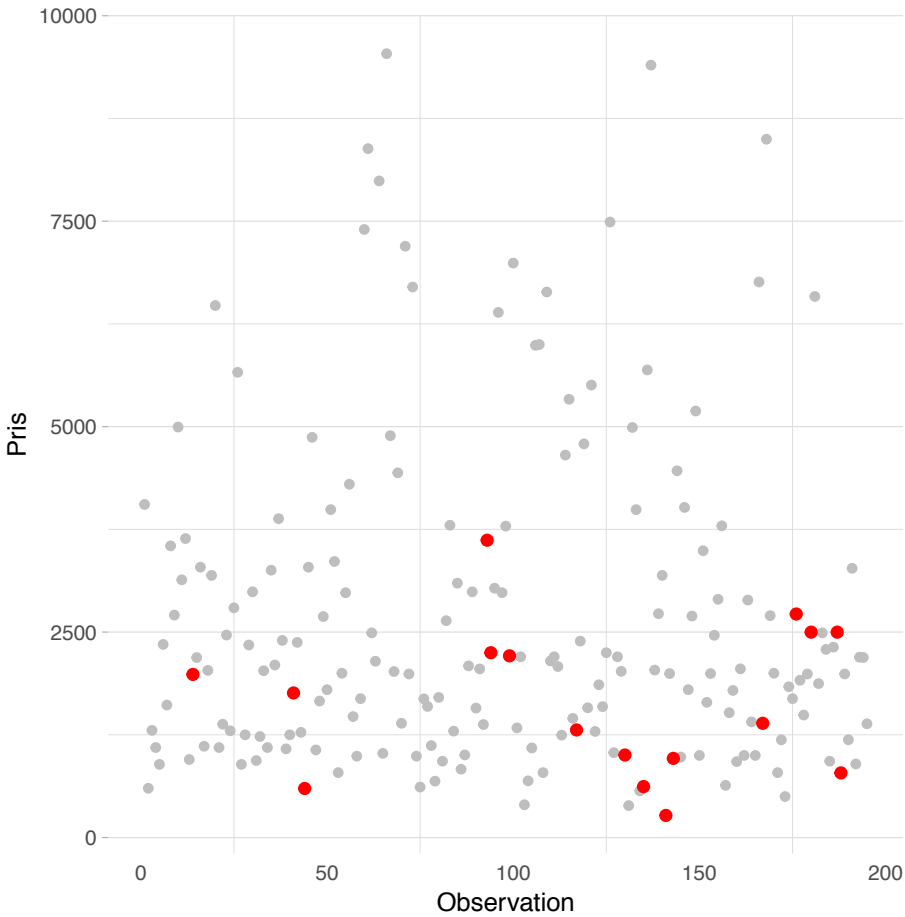
Som tidigare nämnt är slutmodellen bäst lämpad för prediktion. I Figur 8 ser vi att observationerna blir mer utspridda ju dyrare priset är, alltså förekommer större prediktionsfel vid dyrare priser. Värt att nämna är att MSEP kan vara missvisande om framtida observationer skiljer sig mycket från datamaterialet som vi använde i regressionsanalysen för att ta fram modellen. Om framtida observationer till exempel har mycket högre eller mycket lägre värden är det möjligt att prediktionsmodellen inte längre är effektiv. Eftersom det inte finns någon exakt gräns på ett bra eller dåligt MSEP-värde är det svårt att säga om prediktionsmodellen är duglig, men det går att konstatera att reducerad modell 3 är en bättre prediktionsmodell än reducerad modell 1 och ju lägre MSEP-värde, desto bättre prediktion.

Det finns såklart fler variabler som kan ha en inverkan på priset, men som inte togs med i datamängen. Framförallt användarvänligheten kan tänkas ha en inverkan och mobiltelefonens funktioner som till exempel stegräknare och olika kamerafunktioner osv. Dessa har inte tagits med då de inte varit tillgängliga på specifikationssidorna.

Som vi tidigare nämnt har data samlats in manuellt, där vi slumpat fram 200 observationer ur 1013 st på prisjakt. Vi har sedan tagit bort några observationer på grund av saknade värden och dubletter och även lagt undan observationer för prediktionssyfte. Detta har då lett till att vi haft totalt 131 observationer att arbeta med, vilket är 12.9% av det som fanns på prisjakt. Slutresultatet kan ha påverkats av detta och kan ge ett missvisande resultat. Det bästa skulle vart om vi kunde tagit med alla observationer som fanns på prisjakt för att få bättre skattningar, men på grund av tidsbrist har det inte varit möjligt. Då vi har samlat in data manuellt finns det risk för felinmatad information. Då en vill göra en liknande analys med minimal osäkerhet bör denne ha med fler variabler och betydligt fler observationer. Automatisk nedladdning av datamaterial är även till fördel för att spara på tid och

mänskliga felinmatningar. Det kan även finnas en viss osäkerhet på prisjakts hemsida. Vissa hemsidor kanske hade utförsäljning eller kampanjpriser just den dagen vi samlade in data, då har vi alltså möjligtvis fått med en eller flera observationer med ett pris som är mycket lägre än vad den borde vara.

A Datamaterial



Figur 9: Observationer mot priser. De rödmarkerade är observationer som saknar värden.

B Multikollinjäritet

Variabel	VIF
DTillverkare	2.080011
Skärmstorlek	43.922301
Skärmupplösning	2.849883
InterntMinne	3.729101
Lansering	2.952699
ProcessorTyp	3.623210
Bakkamera	4.410304
Framkamera	2.852136
RAM	7.533564
Minneskort	1.711143
Vattentålig	1.837772
NFC	1.883819
Fingeravtryck	3.090695
Längd	42.467733
Bredd	31.176090
Djup	5.341990
Vikt	5.642736
Batterikapacitet	4.661347
UtbytbartBatteri	1.851386
Dglas	22.279851
Dmetall	28.944078
Dplast	37.400295

Tabell 13: VIF-värden på icke-transformerade förklarande variabler.

Variabel	VIF
DTillverkare	1.978295
InterntMinne	2.442338
Lansering	2.553880
ProcessorTyp	2.997700
Bakkamera	4.061860
Framkamera	2.758440
Minneskort	1.600750
Vattentålig	1.572455
NFC	1.749417
Fingeravtryck	2.694900
Vikt	3.695133
Batterikapacitet	3.906245
UtbytbartBatteri	1.624170
Dglas	22.031769
Dmetall	28.620883
Dplast	37.109238
Volym	3.387271
Skärm	2.520410

Tabell 14: VIF-värden på transformerade förklarande variabler.

Referenser

[1] Andersson, R. & Tyrcha, J. (2015). *Notes in Econometrics.* Stockholm: Department of Mathematics, Stockholm University.

[2] Compare before buying. (2018). Tillgänglig på `https://www.comparebeforebuying.com`. (Hämtad 28-01-18).

[3] Davidsson, P & Findahl, O. (2016). *Svenskarna och internet 2016.* Stockholm: Internetstiftelsen i Sverige. Tillgänglig på: `https://www.iis.se/docs/Svenskarna_och_internet_2016.pdf`. (Hämtad 25-03-18).

[4] DeviceSpecifications. (2018). Tillgänglig på `https://www.devicespecifications.com`. (Hämtad 28-01-18).

[5] GSMArena. (2018). Tillgänglig på `https://www.gsmarena.com`. (Hämtad 28-01-18).

[6] GSMchoice. (2018). Tillgänglig på `http://www.gsmchoice.com/en/`. (Hämtad 28-01-18).

[7] Held, L. & Sabanés Bové, D. (2014). *Applied Statistical Inference.* Berlin: Springer-Verlag Berlin Heidelberg.

[8] Nationalencyklopedin. *aluminium.* NE Nationalencyklopedin AB. Tillgänglig på: `https://www.ne.se/uppslagsverk/encyklopedi/lang/aluminium`. (Hämtad 05-04-18).

[9] Plastbearbetning. (2015). *Polykarbonat.* Plastbearbetning AB. Tillgänglig på: `https://www.plastbearbetning.se/polycarbonat/`. (Hämtad 05-04-18).

[10] Prisjakt Sverige AB. (2018). Tillgänglig på: `https://www.prisjakt.nu/kategori.php?l=v1436`. (Hämtad 28-01-18).

[11] Sundberg, R. (2014). *Lineära Statistiska Modeller* - 2nd ed., Stockholm: Department of Mathematics, Stockholm University.