PM till Nämnden för KPI
Sammanträde nr 5

PMU/MFS
Katarina Lashgari
ES/PR
Alexandra Garcia Nilsson
John Johansson
Sofie Öhman

2018-10-16

Hedonisk kvalitetsvärdering av hemelektronik med webbskrapning

För diskussion och information

Enheten för prisstatistik fortsätter utforska möjligheter att tillämpa hedoniska priskvalitetsvärderingar på olika produktgrupper, baserat på data insamlad med webbskrapning från en webbutik.

Rapporten presenterar en ny och fördjupad analys av statistiska modeller för hedonisk kvalitetsvärdering av kaffebryggare samt en pilotstudie för hedoniska modeller gällande diskmaskiner. Diskussion om vilken modell som bör väljas för kvalitetsvärdering av kaffebryggare välkomnas så väl som inlägg kring hantering av data.

INNEHÅLL

1 Bakgrund

Prisenheten har sedan en tid tillbaka drivit ett projekt som syftar till att utveckla hedoniska modeller baserade på webbskrapad data. Tidigare pm till nämnden har presenterat modeller för kaffebryggare och TV-apparater. Vid tidigare studier har stort fokus lagts vid att utveckla kod för webbskrapning av priser och egenskaper och mindre utrymme har funnits för att gå på djupet avseende de hedoniska modellerna. När vi jobbat vidare med modellerna ser vi att det finns potential för ytterligare förbättring och utveckling av dem. Därmed ser vi att det finns anledning att arbeta vidare med kaffebryggare. Kaffebryggare är dessutom en relativt enkel produkt att arbeta med, då den jämfört med andra produkter inom hemelektronik har förhållandevis få egenskaper. Genom att utveckla och förbättra vår hedoniska modell för kaffebryggare kommer vi att få värdefulla kunskaper inför framtida utveckling av hedoniska modeller för mer komplexa produkter.

När analysen av kaffebryggare presenterades vid nämndens möte nr. 4 uppkom synpunkter avseende statistiska aspekter såsom samspel mellan olika variabler, kollinjäritet, proxyvariabler samt olika anpassningsmått, till exempel R^2. Det rådde även delade meningar kring valet att använda trunkerad data.

Erfarenheterna från tidigare studier väcker även intresse för att utveckla modeller baserade på webbskrapad data för andra produkter inom hemelektronik, så som diskmaskiner.

2 Syfte

Syftet med denna studie är att fortsätta utreda hedoniska modeller för kvalitetsvärdering av kaffebryggare, med fördjupad statistisk analys som tar hänsyn till kommentarer från tidigare nämndmöte. Dessutom avses att ta arbetet med hedoniska modeller baserade på webbskrapad data vidare, genom att undersöka möjligheterna att utveckla en modell för diskmaskiner.

Slutligen ämnar studien bredda kunskapen kring de möjligheter och utmaningar som kan uppstå vid användningen av pris- och egenskapsdata insamlad genom webbskrapning för hedoniska modeller.

3 Resultat

Nedan presenteras de resultat som erhållits från studierna av kaffebryggare respektive diskmaskin. Datamaterialet som ligger till grund för de konstruerade modellerna har samlats in från en prisjämförelsesajt under 2018. Hedoniska index har sedan beräknats på den data som använts för att beräkna KPI under 2017.

3.1 Kaffebryggare

Det analyserade datamaterialet innehåller tre kontinuerliga variabler: *Pris*, *Volym* och *Effekt* samt några kvalitativa variabler som beskriver olika egenskaper av en kaffebryggare, nämligen Märke (20 nivåer) och 10 stycken variabler med 2 nivåer var: *Avkalkningsprogram*, *Display*, *Dubbelbryggare*, *Kaffekvarn*, *Termoskanna*, *Timer*, *Signallampa*, *Droppstopp*, *Autoavstängning* och *Värmehållning*. I denna analys, har vi konstruerat och analyserat en ny kvalitativ variabel, som vi kallade *Volym_faktor*. Den innehåller två nivåer, kallade *Liten* och *Sto*r, där *Liten* representerar alla kaffebryggare vars volym är lika med eller mindre än 1,5 liter, medan *Stor* representerar de övriga. Kaffebryggare vars volym är mindre än 0,5 liter har inte analyserats då de inte ingår i prismätningarna för KPI.

Utan missing values (eng), innehåller data 1488 prisobservationer, som svarar mot 207 unika kaffebryggare. Varje unik kaffebryggare har minst två upprepade observationer på *Pris*, som

erbjuds av olika webbnetförsäljare *(Butik)* vid samma tidpunkt (nämligen, 2018-09-24). Från en statistisk synpunkt, är en möjlig nackdel med upprepade observationer att de inte nödvändigtvis är ömsesidigt oberoende. Om analysen visar att så är fallet är en möjlig lösning att analysera medelpriserna för varje unik kaffebryggare.

Variabeln *Butik* kan betraktas som en proxyvariabel för nivån av försäljnings-och reparationsservicen. Men med tanke på att byte sker endast inom en och samma butik känns det irrelevant att formulera statistiska modeller med *Butik* som förklarande variabel.

Vi har valt att avstå från att trunkera data då eliminering av några största och minsta värden inte nödvändigtvis löser problemet med avvikande observationer. Dessutom kan det leda till en viss förlust av information om prisvariation.

3.1.1 Statistisk analys

Fördelningarna av de ovanstående variablerna samt deras relationer till varandra presenteras grafiskt i Figur 1A-5A i Appendix. Baserat på Figur 3A, är det motiverat att logaritmera de kontinuerliga variablerna för att kunna uppfylla antagandet om linjära relationer. På så sätt, får vi en log-log regressionsmodell. Modellen som presenteras nedan har den bästa anpassningen till data i termer av AIK (se Förklaring till Tabell 1):

Statistisk modell 1 (data med 1488 observationer, Märke 1, Märke 3, Märke 9, Märke 13, Märke 17 ingår i basgruppen. Ursprungligen är det Märke 9 som ingick i basgruppen då den innehåller störst antal observationer)

$$\log(Pris) = \alpha + \beta_1 * \log(Effekt) + \beta_2 * \log(Volym) + \sum_i c_i * Märke_i + \gamma_1 * Avkalkningsprogram +$$
$$+ \gamma_2 * Autoavstängning + \gamma_3 * Display + \gamma_4 * Dubbelbryggare + \gamma_5 * Kaffekvarn +$$
$$+ \gamma_6 * Termoskanna + \gamma_7 * Timer + \gamma_8 * Signallampa + \gamma_9 * Varmehallning +$$
$$+ \gamma_{10} * \log(Volym) : Märke8 + \varepsilon, \qquad där\ \epsilon \sim iid\ N(0, \sigma^2)\ .$$

Det numeriska resultatet av anpassningen av *Modell 1* presenteras i *Tabell 1* nedan. Diagnostikplottarna presenteras i Figur 6A i Appendix.

Tabell 1. Resultat av anpassningen av *Modell 1*

	Estimate	Std. Error	t value	Pr(>|t|)	VIF	Omräknade skatt.na
(Intercept)	-1.377	0.315	-4.37	<<0.001	-	0.25 :=e^(-1.38)
log(Effekt)	1.087	0.045	24.10	<<0.001	2.9	1.11 :=1.1^1.087
log(Volym)	-0.241	0.055	-4.37	<<0.001	2.1	0.98 :=1.1^(-0.241)
Märke2	0.334	0.039	8.66	<<0.001	1.4	1.40 :=e^0.334
Märke4	1.557	0.080	19.43	<<0.001	2.4	4.74 på samma sätt
Märke5	0.528	0.081	6.52	<<0.001	1.1	1.70 på samma sätt
Märke6	0.177	0.040	4.39	<<0.001	1.7	1.19 på samma sätt
Märke7	0.249	0.071	3.50	<<0.001	1.1	1.28 på samma sätt
Märke8	0.038	0.087	0.43	0.670	2.1	1.04 på samma sätt
Märke10	0.900	0.034	26.32	<<0.001	2.2	2.46 på samma sätt
Märke11	0.173	0.041	4.23	<<0.001	1.3	1.19 på samma sätt
Märke12	0.126	0.032	3.91	<<0.001	1.4	1.13 på samma sätt
Märke14	0.162	0.043	3.74	<<0.001	1.2	1.18 på samma sätt
Märke15	0.258	0.030	8.70	<<0.001	1.6	1.29 på samma sätt
Märke16	0.191	0.061	3.14	0.002	1.1	1.21 på samma sätt
Märke18	-0.198	0.087	-2.28	0.023	1.1	0.82 på samma sätt
Märke19	0.175	0.064	2.72	0.007	1.1	1.19 på samma sätt
Märke20	0.255	0.061	4.21	<<0.001	1.1	1.29 på samma sätt
Avkalkningsprogram	0.271	0.028	9.85	<<0.001	1.9	1.31 på samma sätt
Auto_avstangning	0.163	0.047	3.44	<<0.001	1.9	1.18 på samma sätt
Display	0.231	0.028	8.23	<<0.001	2.3	1.26 på samma sätt
Dubbelbryggare	0.469	0.037	12.50	<<0.001	1.5	1.60 på samma sätt
Kaffekvarn	0.769	0.045	16.96	<<0.001	1.4	2.16 på samma sätt
Termoskanna	0.184	0.021	8.97	<<0.001	1.5	1.20 på samma sätt
Timer	-0.108	0.022	-4.90	<<0.001	1.7	0.90 på samma sätt
Signallampa	-0.094	0.019	-5.03	<<0.001	1.4	0.91 på samma sätt
Varmehallning	0.051	0.018	2.79	0.005	1.3	1.05 på samma sätt
log(Volym):Märke8	-0.417	0.198	-2.11	0.035	2.2	0.96 :=1.1^(-0.417)
R^2 = 0,80, **R^2_ adjusted** = 0,79, **AIK** = 674						
P-värdena för tre test utförda i syfte att kontrollera modellantagandena						
Shapiro-Wilk normality test: *p*-value < 2e-15 (nullhypotesen om normalitet förkastas)						
Breusch-Pagan (BP) test for heteroscedasticity: *p*-value=0,013 (nullhypotesen om konstanta variansen förkastas)						
Box-Ljung test for independence: *p*-value < 2e-16 (nullhypotesen om oberoendet förkastas)						

Förklaring till Tabell 1:

VIF, *variance inflation factor* (eng), representerar ökningen i varians av en parameter på grund av korrelationen mellan förklarande variabler, dvs. kollinjäritet. VIF-värdet som är större än 5 eller 10 är indikation på att motsvarande koefficient är dåligt skattad på grund av multikollinearitet.

AIK (Akaikes informationskriterium) är ett anpassningsmått som används som ett medel för modellval. AIK är användbart eftersom det bestraffar uttryckligen eventuella överflödiga parametrar i modellen genom att lägga till $2*(p+1)$ till (-2*logLikelihood), där p är antalet parametrar i modellen samt 1 adderas för den beräknade variansen. Bland några modeller väljer man den med det minsta AIK-värdet.

Baserat på $R^2 = 0{,}80$ och $R^2 adjusted = 0{,}79$ har modellen en bra anpassningsgrad till data. Emellertid noterar vi två aspekter. Den första är oväntade tecken för vissa skattade effekter, t.ex. en negativ effekt av *Volym* på *Pris*. Den andra aspekten är att de tre huvudantagandena inte är uppfyllda (se Tabell 1 och Figur 6A i Appendix.). Residualerna är varken normalfördelade eller oberoende sinsemellan samt antagandet om konstant varians är inte heller uppfyllt. Allt detta gör skattningarna otillförlitliga, vilket motiverar en fortsatt analys där responsvariabeln är det logaritmerade medelpriset, *log(Medelpris)*. Den resulterande modellen ges av:

Statistisk modell 2 (data med 207 observationer, *Märke 9* som basmärke):

$$\log(Medelpris) = \alpha + \beta_1 * \log(Effekt) + \sum_i c_i * Märke_i + \gamma_1 * Avkalkningsprogram +$$

$$+ \gamma_2 * Display + \gamma_3 * Dubbelbryggare + \gamma_4 * Kaffekvarn + \gamma_5 * Termoskanna + \varepsilon,$$

$$där\ \epsilon \sim iid\ N(0, \sigma^2)\ .$$

Det numeriska resultatet av anpassningen av *Modell 2* ges i Tabell 2 nedan. Diagnostik-plottarna presenteras i Figur 7A i Appendix.

Tabell 2. Resultat av anpassningen *Modell 2*

	Estimate	Std. Error	t value	Pr(>|t|)	VIF	Omräknade skatt:na
(Intercept)	-0.8592	0.8330	-1.031	0.304		0.424
log(Effekt)	1.0228	0.1182	8.650	<<0.001	2.71	1.102
Märke1	0.0857	0.1048	0.818	0.415	1.37	1.090
Märke2	0.2419	0.1368	1.768	0.0787	1.21	1.274
Märke3	0.0816	0.1494	0.546	0.586	1.21	1.085
Märke4	1.3823	0.1639	8.434	<<0.001	1.73	3.984
Märke5	0.5020	0.2222	2.259	0.0251	1.08	1.652
Märke6	0.1507	0.1139	1.323	0.188	1.37	1.163
Märke7	0.1746	0.1482	1.179	0.240	1.18	1.191
Märke8	0.2918	0.1390	2.099	0.0372	1.24	1.339
Märke10	0.9040	0.0905	9.985	<<0.001	2.33	2.470
Märke11	0.2354	0.1306	1.802	0.0732	1.28	1.265
Märke12	0.1784	0.0932	1.915	0.570	1.58	1.195
Märke13	0.2782	0.1839	1.513	0.132	1.11	1.321
Märke14	0.1410	0.1079	1.307	0.193	1.46	1.151
Märke15	0.1921	0.0844	2.275	0.0241	1.97	1.212
Märke16	0.3384	0.1347	2.513	0.0128	1.17	1.403
Märke17	-0.0718	0.1541	-0.466	0.642	1.28	0.931
Märke18	-0.1990	0.1638	-1.214	0.226	1.16	0.820
Märke19	0.5721	0.1659	3.448	<<0.001	1.19	1.772
Märke20	0.4484	0.1277	3.511	<<0.001	1.22	1.566
Avkalkningsprogram	0.1840	0.0750	2.452	0.0152	1.46	1.202
Display	0.1864	0.0635	2.936	0.00376	1.59	1.205
Dubbelbryggare	0.3312	0.1210	2.738	0.00680	1.24	1.393
Kaffekvarn	0.7687	0.1201	6.402	<<0.001	1.37	2.157
Termoskanna	0.2385	0.0504	4.732	<<0.001	1.22	1.269

R^2 = 0.82, **R^2_adjusted** = 0.80, **AIK** = 116

P-värdena för tre test utförda i syfte att kontrollera modellantagandena

Shapiro-Wilk normality test: *p*-value < 4e-05 (nullhypotesen om normalitet förkastas)
Breusch-Pagan (BP) test for heteroscedasticity: *p*-value=0,723 (nullhypotesen om konstant varians förkastas ej)
Box-Ljung test for independence: *p*-value = 0.2 (nullhypotesen om oberoendet förkastas ej)

Baserat på $R^2 = 0{,}82$ och R^2 adjusted $= 0{,}80$, kan vi dra slutsatsen att *Modell 2* har en bra anpassningsgrad till data. Vidare, har varje skattad koefficient förväntade tecken samt de omräknade skattningarna verkar ha en meningsfull tolkning. Som exempel kan nämnas att den omräknade skattningen för *Märke 4*, 3,984, indikerar att priset för deras kaffebryggare förväntas vara nästan fyra gånger högre än för kaffebryggare av basmärket *Märke 9* (när *Effekt* och *Volym* hålls konstant). Den här slutsatsen verkar stämma överens med boxplottarna i Figur 3.

Som vi ser i Tabell 2, indikerar VIF- värdena att vi inte har problem med multikollinearitet för *Modell 2*. Det är värt att påpeka att det inte var möjligt att introducera samspel mellan variablerna, i synnerhet mellan *log(Effekt)* och de andra variablerna. Detta eftersom införandet av samspelstermerna ledde till en extremt stark multikollinearitet, vilket i sin tur gör skattningarna väldigt opålitliga.

Givet p-värde = 0,2 för Box-Ljung testet (se Tabell 2), kan vi dra slutsatsen att antagandet om oberoende observationer (eller ekvivalent, residualer) är uppfyllt. Däremot, är antagandet om normalfördelning fortfarande inte uppfyllt, men som känt förväntas fördelningen av de

skattade regressionskoefficienterna gå mot normalfördelning på grund av Centrala gränsvärdesatsen även om residualerna inte är normalfördelade. Därför kan vi dra slutsatsen att *Modell 2* är rimlig och kan accepteras som en slutgiltig modell.

Från KPI-synvinkeln kan *Modell 2*, som innehåller variabler med icke-signifikant effekt, motiveras med att det kan ha en viss effekt på det justerade priset när en kaffebryggare av $Märke_i$ bytts mot en annan kaffebryggare av $Märke_j$ även om de motsvarande regressionskoefficienterna c_i och c_j inte är signifikanta.

För att få djupare insikter om detta, analyserar vi några enklare regressionsmodeller med färre parametrar. Från KPIs perspektiv, är ett möjligt sätt att evaluera modellernas prestanda att jämföra indexserier som baseras på dessa modeller. Statistiskt, kan modellerna jämföras i termer av anpassnings- och prediktionsmått.

För att formulera enklare modeller, modifierar vi *Modell 2* antingen genom att successivt eliminera variabler med icke-signifikant effekt eller genom att gruppera märkena på något sätt. Successiv eliminering av icke-signifikanta variabler från *Modell 2* har lett till *Modell 3* som följer:

Statistisk modell 3 (data med 207 observationer. *Märke9* och alla andra märken som inte är inkluderade i modellen utgör basgruppen)

$$\log(Medelpris) = \alpha + \beta_1 * \log(Effekt) + c_1 * Märke4 + c_3 * Märke10 + c_4 * Märke18 + c_5 * Märke19 + c_6 * Märke20 + \gamma_1 * Display + \gamma_2 * Dubbelbryggare + \gamma_3 * Kaffekvarn + \gamma_4 * Termoskanna + \epsilon, \text{ där } \epsilon \sim iid\ N(0, \sigma^2)$$

Det numeriska resultatet av anpassningen av *Modell 3* ges i Tabell 3 nedan. Diagnostikplottarna presenteras i Figur 8A i Appendix.

Tabell 3. Resultat av anpassningen av *Modell 3*

	Estimate	Std. Error	t value	Pr(>|t|)	VIF	Omräknade skatt:na
(Intercept)	-1.114	0.7537	-1.48	0.141		0.328
log(Effekt)	1.083	0.1083	10.00	<<0.001	2.18	1.109
Märke4	1.170	0.1575	7.43	<<0.001	1.54	3.223
Märke10	0.725	0.0784	9.25	<<0.001	1.68	2.065
Märke18	-0.361	0.1571	-2.30	0.0227	1.03	0.697
Märke19	0.476	0.1627	2.93	0.00383	1.10	1.610
Märke20	0.299	0.1204	2.49	0.0137	1.04	1.349
Display	0.217	0.0583	3.71	<<0.001	1.30	1.242
Dubbelbryggare	0.345	0.1180	2.92	0.00389	1.14	1.412
Kaffekvarn	0.744	0.1155	6.44	<<0.001	1.22	2.104
Termoskanna	0.243	0.0491	4.94	<<0.001	1.11	1.275

R^2 = 0.80, **R^2_ adjusted** = 0.79, **AIK** = 111

P-värdena för tre test utförda i syfte att kontrollera modellantagandena

Shapiro-Wilk normality test:	*p*-value < 3.0e-05	(nullhypotesen om normalitet förkastas)
Breusch-Pagan (BP) test for heteroscedasticity:	*p*-value=0,342	(nullhypotesen om konstant varians förkastas ej)
Box-Ljung test for independence:	*p*-value = 0.4	(nullhypotesen om oberoendet förkastas ej)

Som man kan se i Tabell 3, beskriver *Modell 3* prisvariationen nästan lika bra som *Modell 2*: R^2 minskas obetydligt från 0,82, till 0,80; $R^2 adjusted$ blir 0,79 istället för 0,80, vilket också är en försumbar minskning.

Notera också att AIK-värdet minskas från 116 till 111, vilket tyder på en bättre anpassningsgrad. Vidare, indikerat av VIF- värdena i Tabell 3, är graden av multikollinjäritet för *Modell* 3 acceptabel. De omräknade skattningarna pekar inte heller på några problem med tolkningen. Allt detta tillsammans föreslår att den enklare modellen, dvs. *Modell 3*, är att föredra (åtminstone från den statistiska synpunkten).

Ett annat sätt att modifiera *Modell 2* är att gruppera märkena på ett tolkbart sätt, något som gjordes i tidigare analyser. Till exempel, kan man slå ihop *Märke* 4 med *Märke* 10. Detta kan motiveras med faktumet att dessa två märken antas att producera kaffebryggare av lyxkvalitet, vilket reflekteras i högre priser. Den resulterade regressionsmodellen, där den nya gruppen betecknas med *Märkena_4_10*, ges av *Modell 4* nedan:

Statistisk modell 4 (data med 207 observationer. Basgruppen innehåller *Märke 9* och alla andra 14 märken som ej är inkluderade i modellen)

$$\log(Pris\_medel) = \alpha + \beta_1 * \log(Effekt) + c_1 * Märke\_4\_10 + c_3 * Märke18 + c_3 *$$

$$Märke19 + c_4 * Märke20 + \gamma_2 * Display + \gamma_3 * Dubbelbryggare + \gamma_4 * Kaffekvarn + \gamma_5 *$$

$$Termoskanna + \epsilon, \text{ där } \epsilon \sim iidN(0, \sigma^2)$$

Det numeriska resultatet av anpassningen av *Modell 4* ges i Tabell 4. Diagnostikplottarna presenteras i Figur 9A i Appendix.

Tabell 4. Resultat av anpassningen av *Modell 4*

	Estimate	Std. Error	t value	Pr(>|t|)	VIF	Omräknade skatt:na
(Intercept)	-1.810	0.7327	-2.47	0.0144		0.164
log(Effekt)	1.182	0.1054	11.21	<<0.001	1.99	1.119
Märke_4_10	0.755	0.0794	9.50	<<0.001	1.92	2.127
Märke18	-0.358	0.1603	-2.23	0.0267	1.03	0.699
Märke19	0.435	0.1654	2.63	0.00922	1.10	1.545
Märke20	0.278	0.1226	2.27	0.0245	1.04	1.321
Display	0.217	0.0595	3.64	<<0.001	1.30	1.242
Dubbelbryggare	0.281	0.1185	2.37	0.0189	1.10	1.324
Kaffekvarn	0.750	0.1178	6.37	<<0.001	1.22	2.118
Termoskanna	0.276	0.0488	5.65	<<0.001	1.06	1.318

R^2 = 0.79, **R^2_ adjusted** = 0.78, **AIK** = 118

P-värdena för tre test utförda i syfte att kontrollera modellantagandena

Shapiro-Wilk normality test: *p*-value < 5e-05 (nullhypotesen om normalitet förkastas)
Breusch-Pagan (BP) test for heteroscedasticity: *p*-value=0,804 (nullhypotesen om konstant varians förkastas ej)
Box-Ljung test for independence: *p*-value = 0.6 (nullhypotesen om oberoendet förkastas ej)

Som vi ser, är *Modell 4* enklare än *Modell 3* i fråga om antalet parametrar. Efter de två ovannämnda märkena slagits ihop i en grupp, blev effekten av vissa variabler icke-signifikant, vilket motiverade deras eliminering från modellen. Detta ledde till en ytterligare reducering av antalet parametrar. Trots detta minskas R^2 och $R^2 adjusted$ obetydligt, jämfört med *Modell 3*. Följaktligen, baserat på R^2 och $R^2 adjusted$, kan vi dra slutsatsen att *Modell 4* passar data nästan lika bra som *Modell 3*.

I Tabell 4 noterar vi också att samtliga VIF- värdena tyder på en acceptabel grad av multikollinearitet för *Modell 4* eftersom alla VIF-värdena är mindre än 5. Även de omräknade skattningarna tycks att ha rimliga tolkningar.

Men AIK-värdet i Tabell 4 säger oss att *Modell 4* passar sämre till data än *Modell 3* eftersom AIK-värdet för *Modell 4* är större än för *Modell 3* (jämför 111 till 118). Statistiskt, kan denna försämring i anpassningsgraden förklaras av faktumet att den skattade koefficienten för *Märke 4* i *Modell 3* skiljer sig signifikant från den skattade koefficienten för *Märke 10* (nullhypotesen att $c_1 = c_3$ förkastades på alla vedertagna signifikansnivåer).

Validering

Ett ytterligare sätt att utvärdera de formulerade modellerna är att jämföra deras förmåga att prediktera. För detta ändamål, användes ett nytt datamaterial, insamlat 2018-10-01, som valideringsdata. Det nya datamaterialet innehåller samma variabler som det analyserade datamaterialet, men antalet observationer är N=205 istället för N=207. Som prediktionsmått användes Mean Squared Error of Prediction (MSEP) , som ges av

$$MSEP = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2.$$

Om det finns några dataset, insamlade vid olika tidpunkter, kan man beräkna MSEP för en given statistisk modell för varje dataset. Genom att studera variationen i MSEPs värden kan man skaffa en viss uppfattning om huruvida nya observationer skiljer sig markant från de ursprungliga observationerna (för minst en variabel). Om det är fallet, är det sannolikt att den statistiska modellen av intresse inte passar till data bra, vilket resulterar i ett större MSEP-värde.

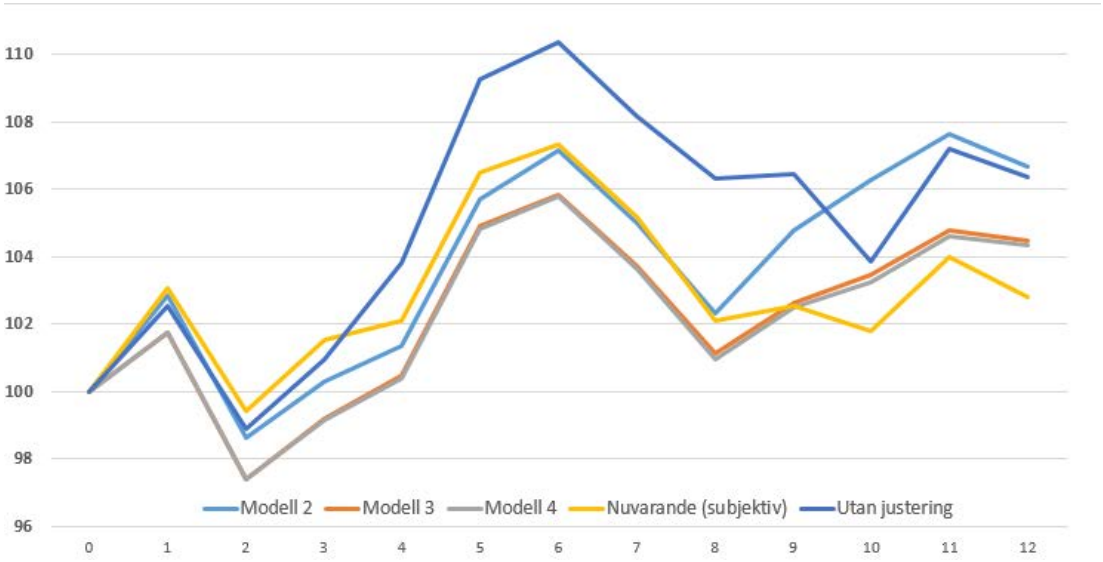
Som sagt, i denna analys finns det ett nytt dataset tillgängligt. Beräkningarna resulterade i MSEP=0,08 för *Modell 2* och MSEP=0,09 för *Modell 3*. Med tanke på att *Modell 3* innehåller 15 parametrar mindre än *Modell 2* är skillnaden i MSEPs värden inte avsevärd. För *Modell 4*, visade sig MSEP vara högre än för *Modell 3*, nämligen 0.12, fastän båda modellerna är nästan ekvivalenta i fråga om antalet parametrar (*Modell 4* har bara 1 parameter mindre än *Modell 3*). Det här resultatet indikerar att prediktionsförmåga av *Modell 4* är lägre än för *Modell 2* och *Modell 3*.

Innan vi diskuterar vilken statistisk modell av de presenterade som kan väljas som en slutgiltig modell, undersöker vi utvecklingen av prisindexet beräknat enligt varje statistisk modell. Notera att vi utesluter *Modell 1* som en rimlig modell från den här undersökningen eftersom modellen har otolkbara och missvisande skattningar. De återstående tre modellerna har demonstrerat en acceptabel anpassningsgrad till data och har en adekvat tolkning. Därför beräknas prisindex baserat på *Modell 2, Modell 3* och *Modell 4*.

3.1.2 Index för modellerna

Figur 1 nedan visar index för kaffebryggare baserat på de hedoniska modellerna 2, 3 och 4 samt direkt jämförelse och subjektiv bedömning. Vi väljer att inte jämföra med den hedoniska modell som presenterats i tidigare pm, då den bygger på trunkerad data och inte är jämförbar med de modeller som nu utvecklats.

Samtliga index med hedonisk kvalitetsjustering och index med subjektiv kvalitetsjustering följer varandra relativt väl fram till augusti månad, under år 2017. Från augusti till september syns en marginell uppgång av index vid subjektiv bedömning, medan uppgången för samtliga index som bygger på hedonisk kvalitetsvärdering har en betydligt brantare ökning. Speciellt syns en uppgång för Modell 2, som fortsätter till december då index istället sjunker mer än övriga hedoniska index. Modell 2 slutar på en betydligt högre nivå än övriga. Index för modellen baserad på subjektiv bedömning avviker främst från de övriga mellan september och oktober. Då sker en minskning för detta index medan vi för de övriga kan utläsa en ökning.



Figur 1. Index enligt de tre godtagbara hedoniska modellerna samt för nuvarande metod och utan kvalitetsjustering.

I Figur 2 nedan är bytena presenterade i mer detalj. När vi studerar *Modell 3* och *Modell 4*, som följer varandra extremt väl för hela året, ser vi att inga byten skett för egenskaper där modellerna skiljer sig åt i hög grad. Exempelvis finns Märke 4 inte med i det material som index bygger på och det har inte skett några byten mellan Märke 10 och annat märke. Att dessa slagits ihop till en gemensam variabel ser vi därav ingen effekt av på index.

Vi studerar *Modell 2* och *Modell 3* i mer detalj vid period 1, 9, 10 och 12, då de främst skiljer sig åt. *Modell 4* är nästan identisk med *Modell 3* i fråga om antalet parametrar och de involverade variablerna. Närmare bestämt, modellerna innehåller sju gemensamma parametrar vars skattningarna är approximativt samma oavsett modell. Således, om byten involverar dessa gemensamma parametrar (eller några av dem), förväntas index för dessa två modeller vara ungefär lika. Som vi ser i Figur 2 är det fallet, vilket innebär att slutsatserna om indexkänsligheten för *Modell 3* gäller även för *Modell 4*.

I period 1 har vi tre byten. Vid det första sker ett byte från Märke 12 till Märke 6, display tillkommer och effekt ökar med 80 watt. Vid andra bytet minskar effekten med 60 watt och avkalkningsprogram försvinner. Vid det tredje bytet ökar effekten med 90 watt. Medan samtliga av dessa förändringarna i egenskaper påverkar index för *Modell 2* ger bytet av märke och avkalkningsprogram ingen påverkan på index för *Modell 3*, då dessa variabler där inte är inkluderade.

I period 9 sker ett byte från en kaffemaskin av Märke 1 till Märke 9, effekten minskar från 1160 till 1050 och varken display eller avkalkningsprogram finns hos maskinen efter bytet. Samtliga av dessa egenskapsförändringar har inflytande på index för *Modell 2* men bara effektförändringen påverkar index för *Modell 3*.

I period 10 sker ett byte från en kaffemaskin av Märke 2 med en effekt på 1100 watt till en kaffebryggare av Märke 9 och en effekt på 1000 watt, där endast effektförändringen påverkar index för Modell 3. Dessutom sker ett byte som inte påverkar kvalitetsvärderingen för någon av modellerna då ingen av egenskaperna ändras, vilket inte påverkar index utöver den rena

prisförändringen. I period 12 sker ett byte från Märke x[1] med display till Märke 11 utan display, vilket inte påverkar index för Modell 3 alls.

Nedgången hos index utan kvalitetsvärdering i period 10 kan förklaras av ett flertal prissänkningar. För de hedoniska modellerna ser vi inte denna sänkning av index, då dessa prissänkningar motverkas av de kvalitetsjusteringar som sker vid bytena under perioden. Hos index utan kvalitetsvärdering däremot har det ena bytet negativ effekt medan det andra inte har någon effekt alls.

Sammanfattningsvis kan de större fluktuationerna av index baserat på *Modell 2* jämfört med de som utläses för *Modell 3* alltså troligtvis förklaras av den högre komplexiteten för *Modell* 2. Denna modell har ju fler parametrar vilket medför att den tar hänsyn till fler egenskapsförändringar, i synnerhet gällande byten av märke. Indexen tycks vara känsliga mot komplexiteten av statistiska modeller som de är baserade på.

Period	Byte		Märke*	Effekt	Display	Termoskanna	Dubbelbryggare	Avkalkningsprogram
	1	Bas	M10	1430	0	0	0	0
			M10	1520	0	0	0	0
1	2	Bas	M12	1000	0	0	0	1
			M6	1080	1	0	0	0
	3	Bas	M1	1160	0	0	0	1
			M1	1100	0	0	0	0
2	4	Bas	M12	1000	0	0	0	0
			M6	1080	1	0	0	0
3	5	Bas	M12	1000	0	0	0	0
			M2	1100	0	0	0	0
	6	Bas	M6	1080	1	0	0	0
			M9	1520	0	0	0	1
	7	Bas	M12	1000	0	0	0	0
4			M9	1520	0	0	1	1
	8	Bas	M12	1000	0	1	0	0
			M12	1000	0	1	0	0
	9	Bas	M10	1430	0	0	0	0
			M10	1520	0	0	0	0
	10	Bas	M6	1080	0	0	0	0
			M9	1080	0	1	0	1
5	11	Bas	M1	1160	0	0	0	1
			M6	1150	1	0	0	0
	12	Bas	M9	1000	0	1	0	0
			M9	1050	0	0	0	0
6	13	Bas	M6	1080	0	0	0	0
			M9	1400	0	0	0	0
	14	Bas	M9	1050	0	0	0	0
			M1	1100	0	0	0	0
8	15	Bas	M9	1155	0	0	0	0
			M9	1000	0	1	0	0
	16	Bas	M11	1050	0	0	0	0
			M12	1000	0	1	0	0
9	17	Bas	M1	1160	1	0	0	1
			M9	1050	0	0	0	0
	18	Bas	M12	1000	0	0	0	0
10			M12	1000	0	0	0	0
	19	Bas	M2	1100	0	0	0	0
			M9	1000	0	0	0	0
12	20	Bas	Mx	1000	1	0	0	0
			M11	1000	0	0	0	0

*Figur 2: Röd markering vid förändrat egenskapsvärde. Grön markering för de egenskaper som endast påverkar index för Modell 2 och blå markering för de som påverkar båda index. *Märke påverkar båda index för modellerna men det finns inget byte av märke som inkluderas av modell 3.*

[1] Märket fanns med i datamaterialet som användes för att bygga modellerna men inte i det som användes för att beräkna KPI under 2017.

3.2 Diskmaskin

3.2.1 Syfte med analysen

Låt oss upprepa att vårt syfte att analysera data för diskmaskiner är att få en preliminär uppfattning om relationer mellan priset och några egenskaper på diskmaskiner. Vi strävar alltså inte att få en slutgiltig modell som kan användas för beräkningar av indexserier.

3.2.2 Beskrivningen av datamaterialet

Det insamlade datamaterialet för diskmaskin har 2423 observationer på 35 variabler, varav 31 beskriver diskmaskinernas olika egenskaper. Det bör också påpekas att data innehåller många missing values, närmare bestämt 1998 observationer. Så vill man analysera de alla 31 egenskaperna, reduceras antalet observationer till 425, vilket motsvarar 80 unika diskmaskiner.

Ytterligare ett problem med datamaterialet är att det verkar innehålla motsägelsefull information. Till exempel att den maximala temperaturen för en viss diskmaskin anges att vara 60^oC, medan det högsta temperaturläget är 75^oC. Den preliminära kontrollen visade dock att det här problemet inte berodde på felinmatning, utan på missvisande information på prisjämförelsesajtens sida.

För att undvika att göra antalet observationer oacceptabelt lågt, fokuserade vi oss på några få variabler som vi tyckte kan ha kapacitet att förklara variationen i priset. Dessa är:

1. log(*Kapacitet*);
2. Den maximala ljudnivå (*maxLjud*), modellerat som en kvalitativ variabel med två nivåer: den lägsta nivån innehåller diskmaskiner vars ljudnivå varierar mellan 39 och 46 dB, medan de övriga diskmaskiner hör till den högsta nivån);
3. Den maximala temperaturen (*maxTemp*), modellerat som en kvalitativ variabel med två nivåer (den lägsta nivån innehåller diskmaskiner vars maximala temperatur är antingen 50^oC , $60^oC, 65^oC$ eller 70^oC, medan den högsta innehåller diskmaskiner vars maximala temperatur är 75^oC;
4. Antal program (AP), modellerat i denna analys som en kontinuerlig variabel;
5. Energiförbrukning per disk (*EPD*), modellerat i denna analys som en kontinuerlig, och
6. *Märke* (17 stycken), modellerat med hjälp av 16 stycken dummyvariabler.

Genom att fokusera på dessa sex variabler reducerades antalet observationer till 832, vilket motsvarar 170 unika diskmaskiner. Precis som i fallet för kaffebryggare analyserade vi det genomsnittliga priset för varje unik diskmaskin för att undvika problem med beroendet mellan residualer. Modellen som vi valde att presentera ges av:

Statistisk modell I:

$$\log(Medelpris) = \alpha + \beta_1 * \log(Kapacitet) + \beta_2 * AP + \beta_3 * EPD + \beta_4 * maxLjudHög + \beta_5 * maxTempHög + \beta_4 * (maxLjudHög : maxTempHög) + \sum c_i * Märke_i + \varepsilon, \text{ där } \varepsilon \sim iid\ N(0,1).$$

Den numeriska resultatet av anpassningen presenteras i Tabell 5 nedan. Diagnostikplottarna presenteras i Figur 10A i Appendix.

Tabell 5. Resultat av anpassningen av *Modell I*

	Estimate	Std. Error	t value	Pr(>|t|)	VIF	Omräkn. skatt:na
Intercept	8.081	0.232	34.9	0.000		3231.81
log(Kapacitet)	0.538	0.124	4.3	<<0.001	3.7	1.05
LjudnivaHög	-0.324	0.055	-5.8	<<0.001	2.4	0.72
MaxTempHög	0.184	0.068	2.7	0.008	1.2	1.20
AP	0.051	0.014	3.7	<<0.001	1.7	1.05
EPD	-0.001	0.000	-4.0	<<0.001	2.5	0.99
Märke 4	-0.612	0.221	-2.8	0.006	1.0	0.54
Märke 5	-0.395	0.117	-3.4	0.001	1.7	0.67
Märke 7	0.116	0.043	2.7	0.008	1.1	1.12
Märke 9	-0.899	0.221	-4.1	<<0.001	1.0	0.41
Märke 10	-0.422	0.157	-2.7	0.008	1.0	0.66
Märke 14	-0.578	0.159	-3.6	<<0.001	1.0	0.56
Märke 16	0.174	0.081	2.2	0.033	1.3	1.19
maxLjudHög:maxTempHög	1.247	0.203	6.2	0.000	1.7	3.48

$R^2 = 0.73,\ \ R^2\ adjusted = 0.71,\ \ AIK = -19$

P-värdena för tre test utförda i syfte att kontrollera modellantagandena

Shapiro-Wilk normality test: *p*-value = 0.74 (nullhypotesen av normalitet förkastas ej)
Breusch-Pagan (BP) test for heteroscedasticity: *p*-value=0,032 (nullhypotesen av konstant varians förkastas)
Box-Ljung test for independence: *p*-value= 0.007 (nullhypotesen av oberoendet förkastas)

Baserat på *p*-värdena för Shapiro-Wilk normality test, Breusch-Pagan test for hetero-scedasticity och Box-Ljung test for independence ser vi att modellen inte uppfyller antagandet om oberoende observationer. Tillsammans med de relativt låga värdena på R^2 och $R^2 adjusted$ kan detta resultat indikera att andra variabler med systematisk inverkan på *Pris* inte är i modellen. Vi ska inte heller glömma att datamaterialet inte är helt pålitligt, vilket gör skattningarna opålitliga. Datamaterialet måste kontrolleras innan en fullständig statistisk analys kan genomföras.

Trots detta osäkra resultat, tycker vi att de valda förklarande variablerna har en god förmåga att förklara variationen i priset och bör definitivt undersökas i framtida analys. Detta motiverar att man först kontrollerar inmatade värden för dessa variabler istället för att kontrollera de alla 35 variablerna. På så sätt kan man spara mycket tid.

4 Diskussion och slutsatser

Gällandande val av den slutgiltiga statistiska modellen för kaffebryggare har analysen visat att *Modell 3* är bäst ur statistisk synpunkt. Modellens påverkan på prisindex visade sig dock vara densamma som påverkan av *Modell 4*, fast *Modell 4* hade sämre anpassning till data. Index likartade utveckling för de båda modellerna kan förklaras av likheten i antalet parametrar och att skattningar för gemensamma parametrar är näst intill lika stora.

Utvecklingen av index för *Modell 2* skiljer ut sig ganska mycket vilket kan förklaras av att modellen tar hänsyn till fler egenskaper. En naturlig fråga är vilken av de tre modellerna som är bäst ur KPIs synvinkel. Det förefaller att vara en svår fråga, vilken vi önskar lyfta till diskussion.

Den andra viktiga aspekten att beakta är de problem som uppstår i samband med datainsamlingen. Beträffande data för kaffebryggare visade det sig att informationen på hemsidan kan uppdateras löpande vilket medförde att vissa observationer i det analyserade datamaterialet var ”felaktiga”. Som tur var hade det inte någon betydande effekt på denna

analys men vi kan inte utesluta att det kan ha större effekt vid andra tillfällen, beroende på information kring vilken variabel som uppdateras.

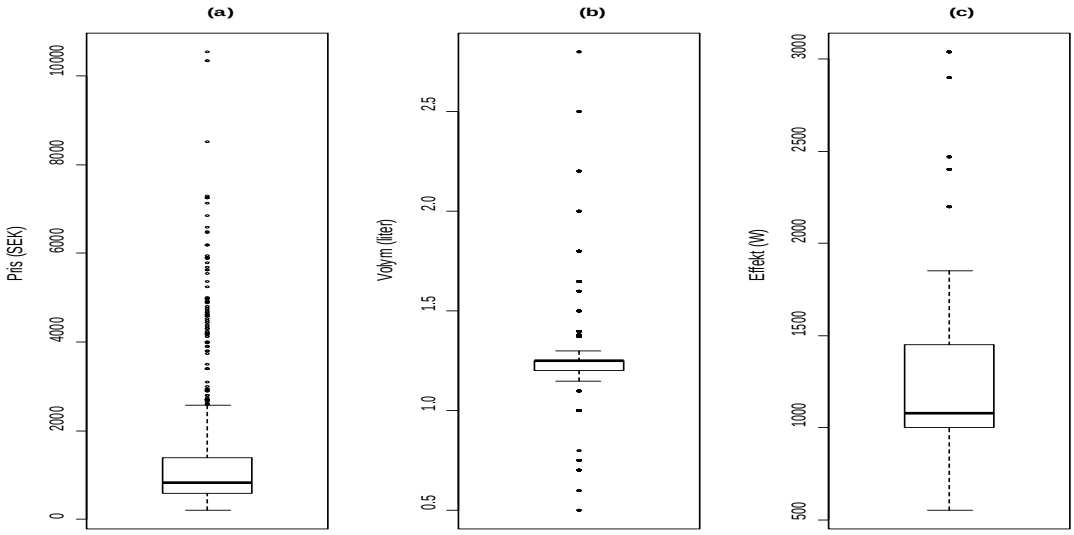
Gällande datamaterialet för diskmaskiner fanns betydligt större problem, bland annat då vi fann att det innehöll en stor andel missing values och motsägande observationer. Detta väcker frågor kring om vår källa är god nog för att samla in egenskapsvärden, eller om vi ska söka information från andra källor.

Lärdomar som har dragits av arbetet är bland annat att variablerna på förhand bör diskuteras igenom innan den statistiska analysen påbörjas. I första hand underlättar det kontroll av datamaterialet; genom att fokusera på några få variabler kan både tid och resurser sparas.

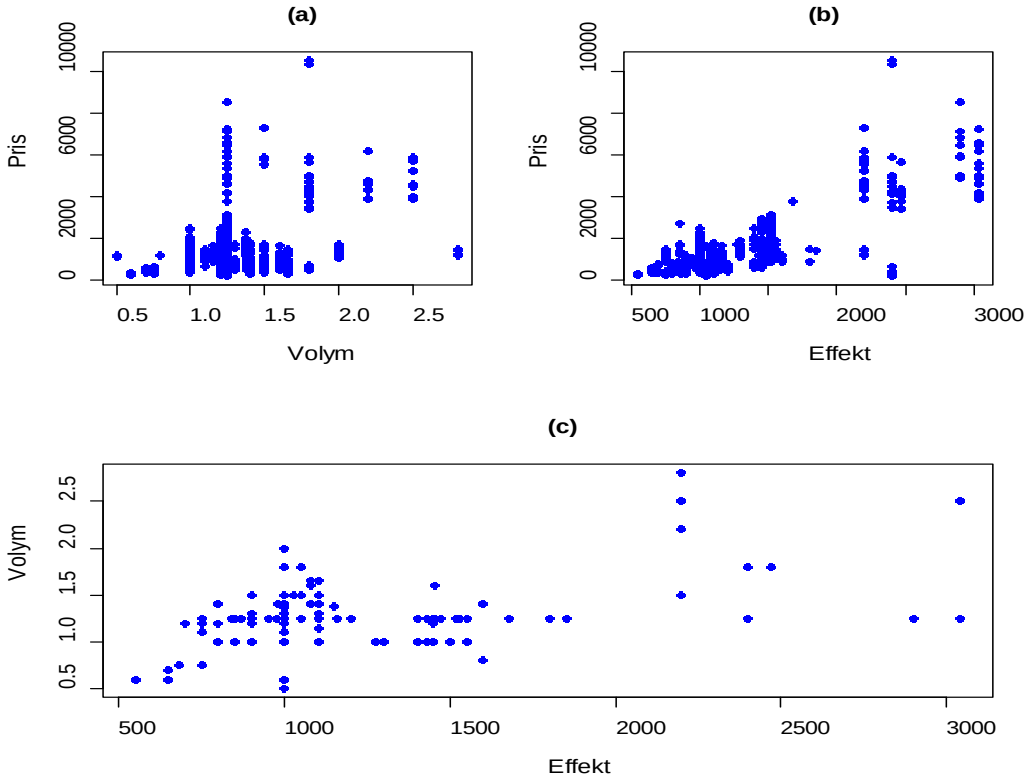
Analysen har även väckt frågor kring rampopulationen, där vi framöver exempelvis funderar kring att ha tydliga beskrivningar av egenskaper för de undersökta varorna. I fallet för kaffebryggare tänker vi bland annat på att exkludera dubbelbryggare och kaffekvarn som egenskaper ur analysen, för att sådana kaffebryggare inte antas vara representativa. Vi har även funderat kring om det vore av intresse att dela in produkterna i fler segment beroende på målgrupp.

Synpunkter gällande vad som bör tas i beaktande vid val av modell, datakällor och datahantering såväl som framtida överväganden välkomnas.

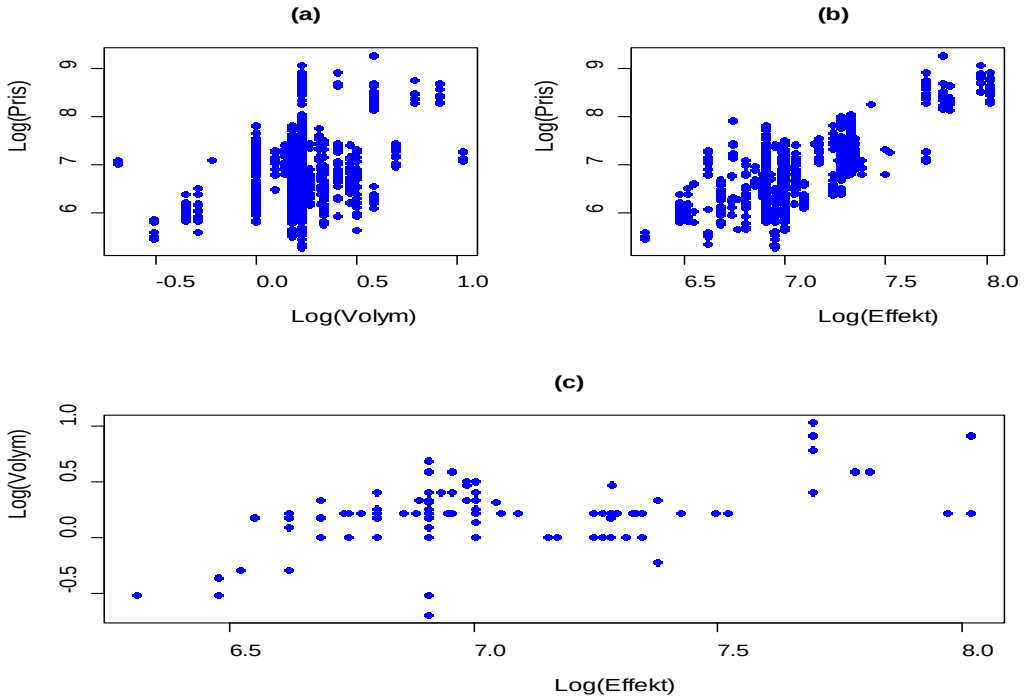
Appendix



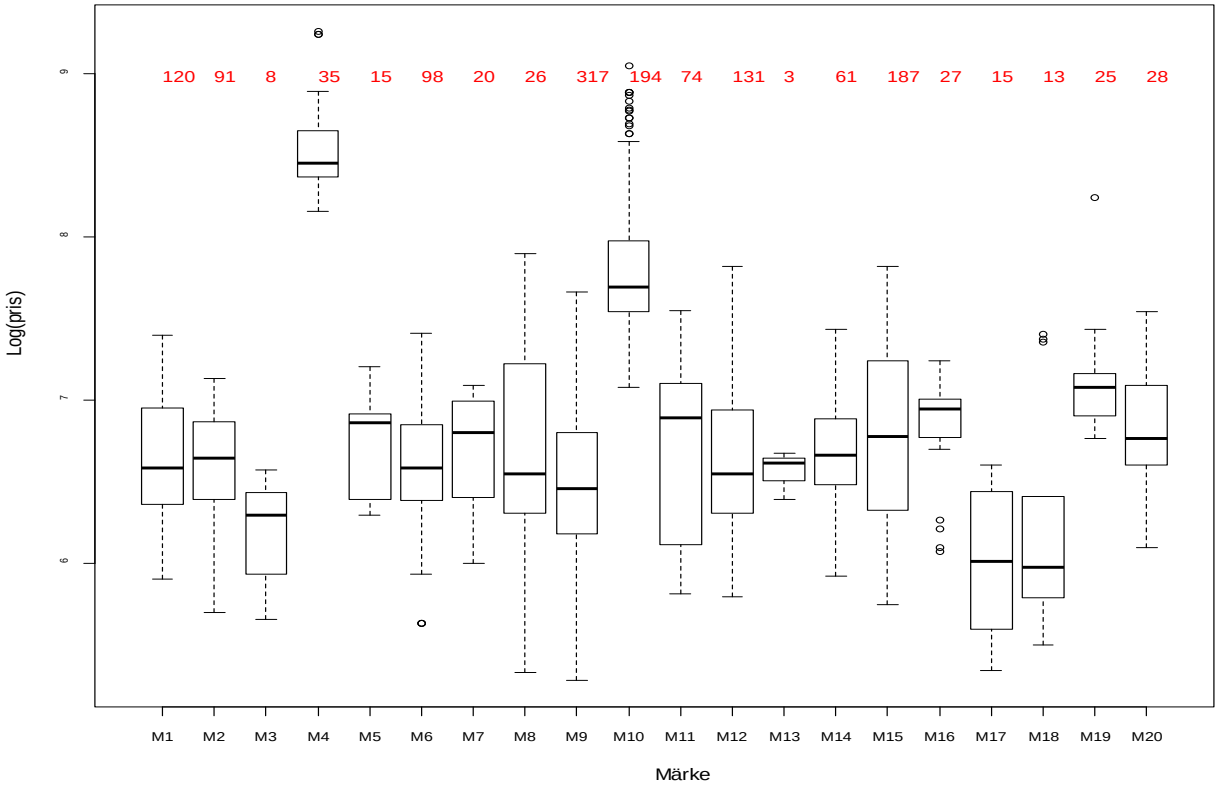
Figur 1A. *Boxplottar för tre kontinuerliga variabler: Pris, Volym, och Effekt. Alla observerade värdena är rimliga.*



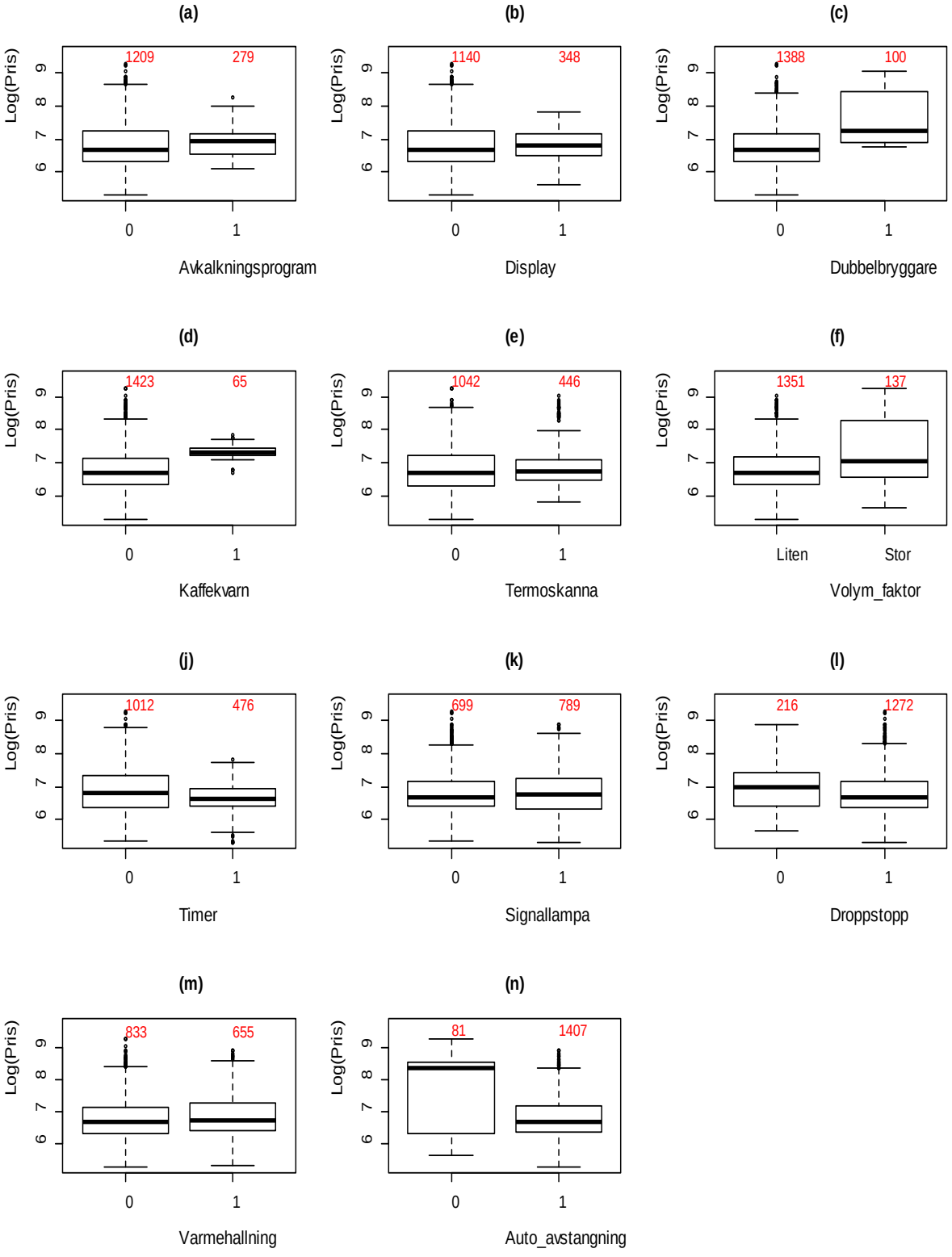
Figur 2A. Scatterplottar för tre kontinuerliga variabler (ursprunglig skala).



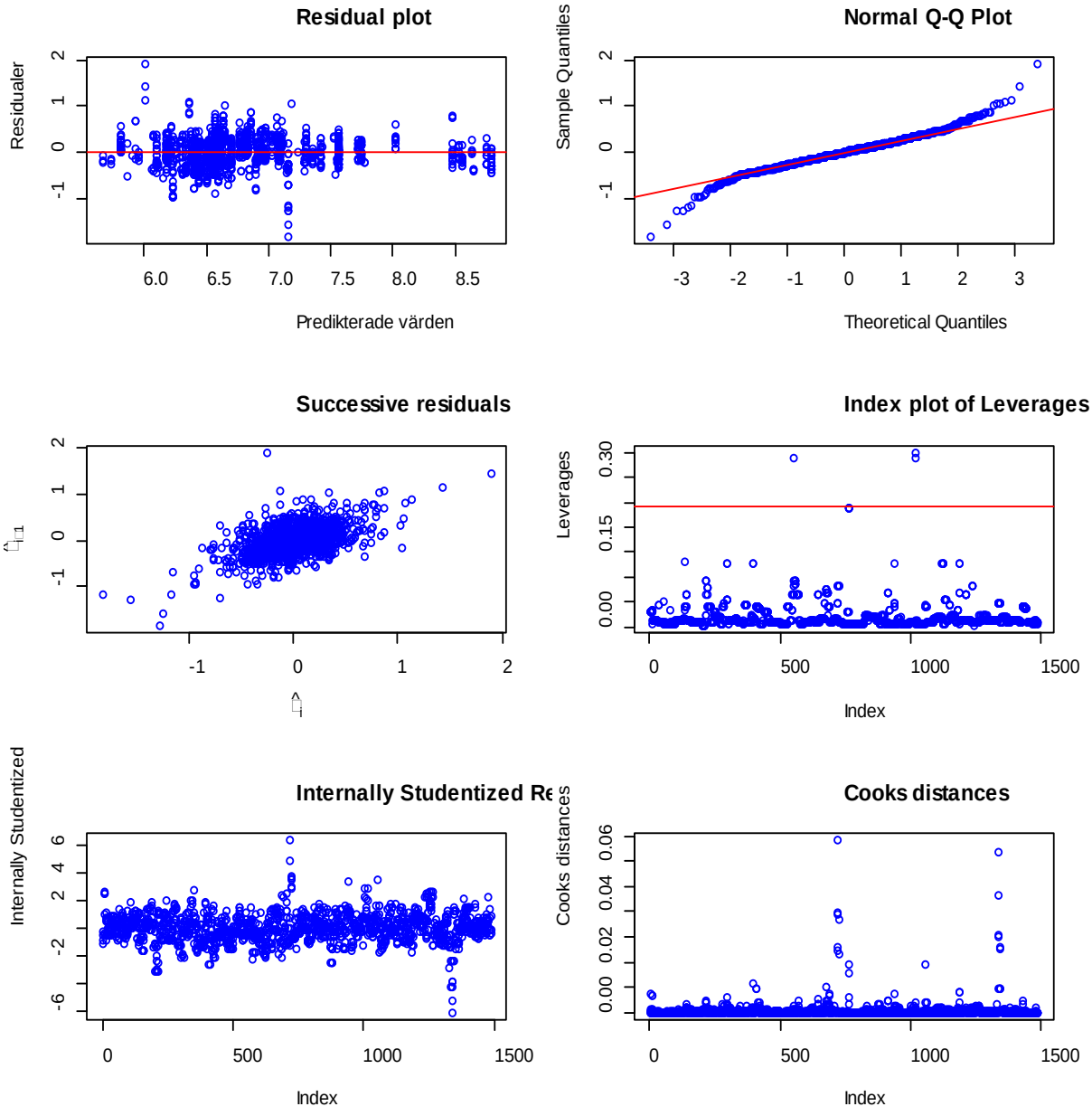
Figur 3A. Scatterplottar för tre kontinuerliga variabler (log-skala).



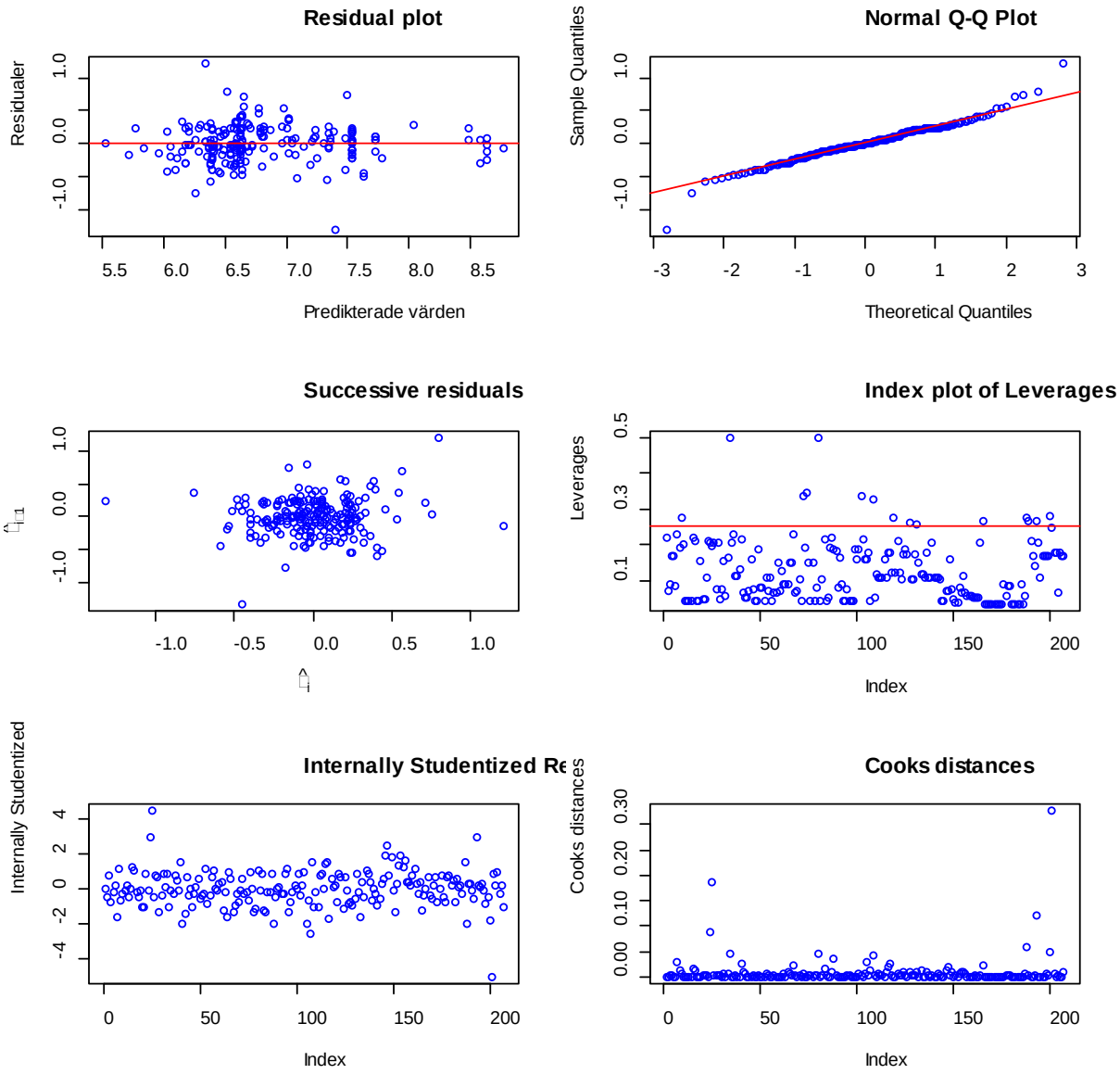
Figur 4A. Fördelningen av *log(Pris)* över variabeln *Märke*. De röda siffrorna i figuren anger antalet observationer (d.v.s.kaffebryggare) för varje märke.



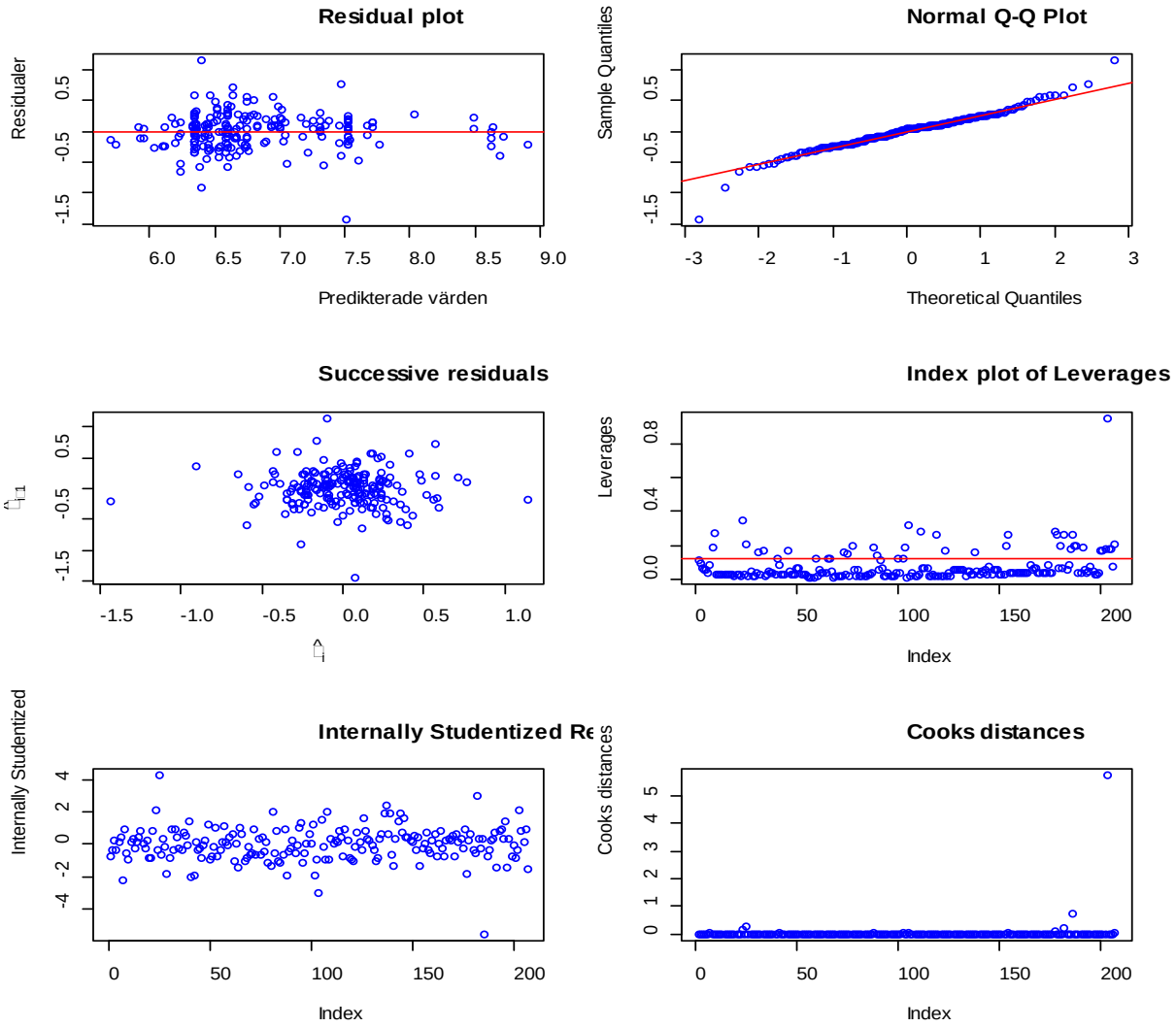
Figur 5A. Fördelningen av *log(Pris)* över 11 kvalitativa variabler. De röda siffrorna i figuren anger antalet observationer för varje nivå av variabeln i fråga.



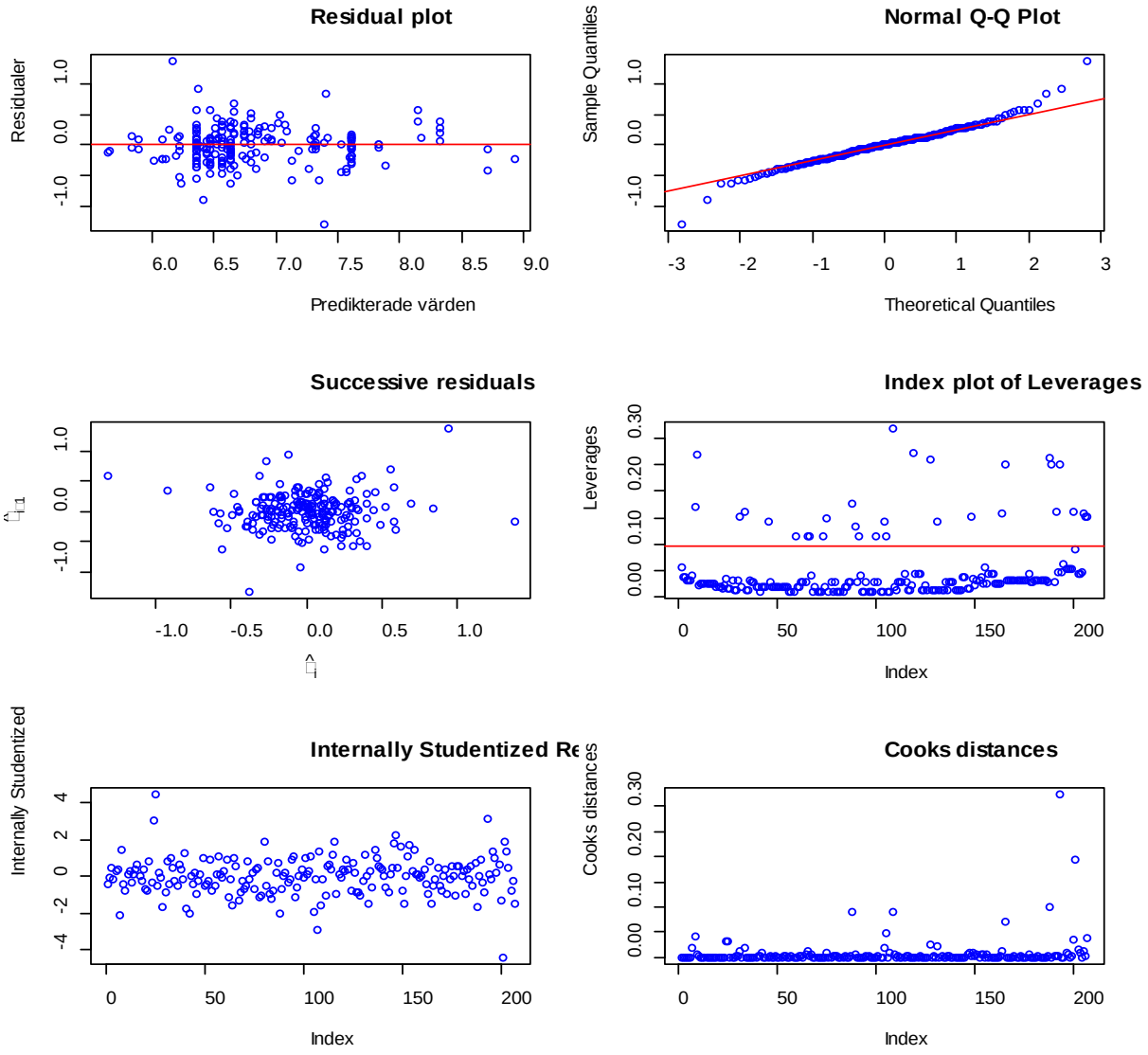
Figur 6A. Diagnostikplottar för *Modell 1*.



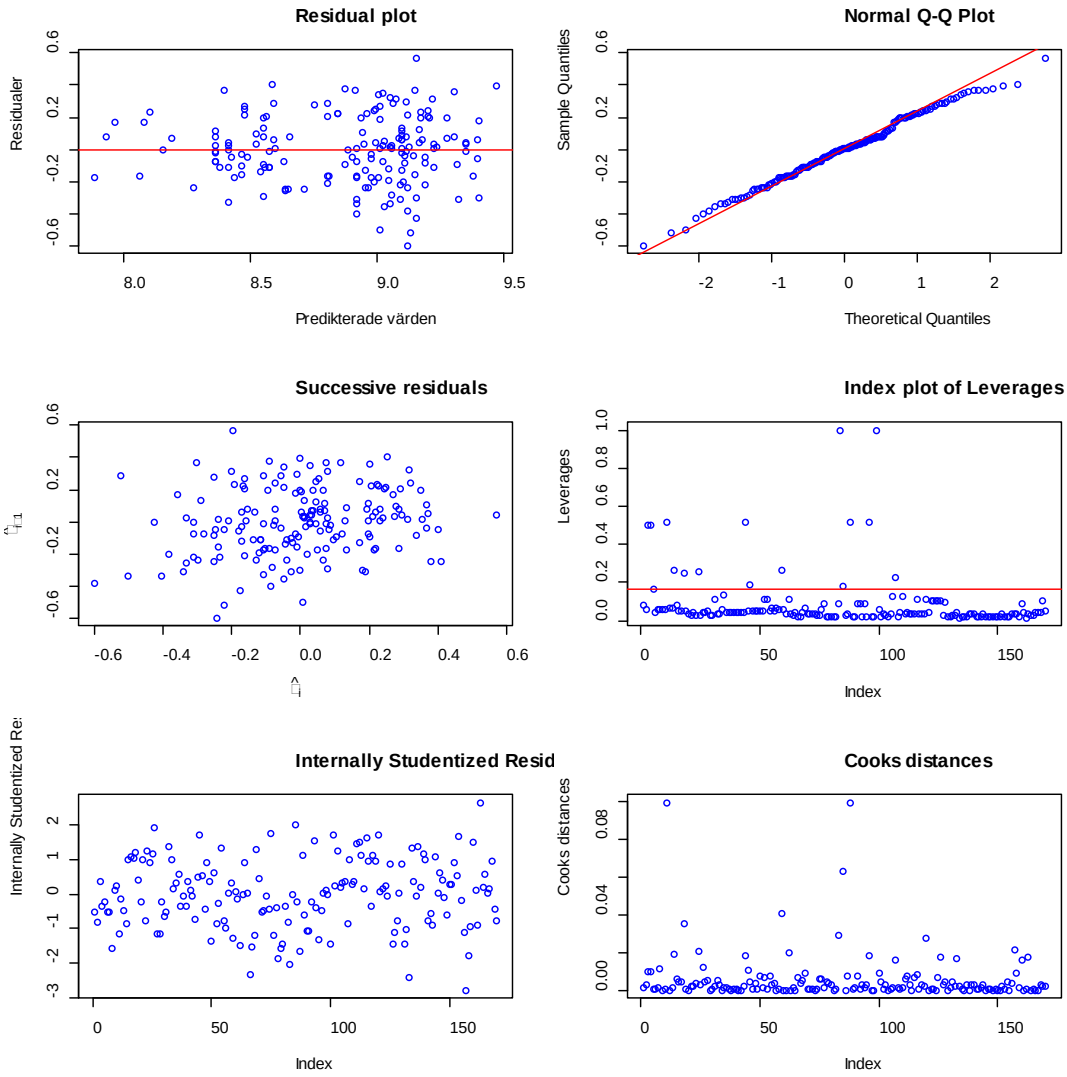
Figur 7A. Diagnostikplottar för *Modell 2*.



Figur 8A. Diagnostikplottar för *Modell 3*.



Figur 9A. Diagnostikplottar för *Modell 4*.



Figur 10A. Diagnostikplottar för *Modell I* (diskmaskinsdata)