Stockholms universitet

Frisörer och Faktorer

Seth Nielsen

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2011:1
Matematisk statistik
Juni 2011

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2011:1**,
http://www.math.su.se/matstat

Frisörer och faktorer

Seth Nielsen*

Juni 2011

Sammanfattning

Rapporten behandlar en mängd insamlad data gällande Stockholm innerstads frisörers omgivning och omständigheter för att avgöra om det finns något mönster i deras prissättning baserat på given information. Uppsatsen går igenom en mängd modeller med eller utan logaritmerat slutpris innan den gemensamt prövar deras förmåga att förutsäga priser utanför underlaget med hjälp av så kallad korsvalidering. Det visar sig att logaritmerade slutpriser inte bidrar till en förbättrad modell i prediktionssyfte. Analysen avslutas med att presentera två modeller som anses likvärdiga statistiskt i sin förmåga att prediktera värden utanför underlaget.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: avatarw@hotmail.com. Handledare: Anders Björkström.

ABSTRACT

This paper analyses an amount of data regarding hair salons collected in the urban part of Stockholm. The analysis controls for a number of information regarding the salons location as well as other circumstances of interest. The paper examines a number of models, both with as well as without the response variable logarithmed, before going on to testing their ability to predict observations outside of their given material for parameter estimation using cross-validation. It becomes evident that the models with a logarithmed response variable do not, in this case, predict values better than other models.

The analysis ends by presenting two models which should be considered equally statistically valid in their ability to predict values outside of their given material for parameter estimation.

INNEHÅLL

1. INLEDNING

Då jag under flera år fått håret klippt i olika lokaler och förstrött försökt urskilja någon form av mönster eller förutsägbarhet utan större lycka har uppsatsämnet alltid funnits i åtanke. Efter att ha gått tre år på Matematik-Ekonomi linjen kändes det dock som att rätt verktyg för att få klarhet i ämnet hade tilldelats.

Förhoppningen var och är att det eventuella mönster som skulle hittas skulle kunna användas i praktiken och ha någon form av värde både för alla de frisörer som självständigt väljer sina priser, men även de som känner att omgivningen i någon utsträckning väljer åt dem.

2. BESKRIVNING AV DATA

Data samlades in under circa 4 veckor där urvalet av salonger skedde slumpmässigt med hjälp av eniro.se under kravet ”Innerstaden Stockholm” där slumpmässigheten säkerställdes genom att använda en slumptalsgenerator vars utfall inte kunde överstiga antalet sökträffar. En annan effekt av att använda eniro.se blev att endast skattebetalande salonger finns med i urvalet.

Insamlingen skedde med hjälp av mig, utskrivna frågeformulär som går att hitta bland bilagorna, en penna och en hel del tålamod och ihärdighet. Särskild vikt lades vid att gå tillbaka till salonger som var för upptagna vid de första besöken så att urvalet inte skulle bli skevt med endast mindre upptagna salonger. Totalt samlades 21 stycken observationer in.

Frågeformuläret reviderades två gånger efter de första dagarna, första gången för lägga till frågor frisörerna själva kom på vilka bedömdes möjliga att samla in på ett sådant sätt att det gick att arbeta med dem, och andra gången för att ändra utformningen på avståndsfrågorna så att personliga tolkningar eliminerades. Endast den slutliga versionen finns bland bilagorna. Samtliga data insamlade innan revideringen har kontrollerats och kompletterats.

Nedan följer en tabell över de insamlade variablerna med tillhörande beskrivning.

Förklarande variabel	Beskrivning
y	Pris för en enkel klippning
x1	Avstånd till närmaste salong som anses konkurrera
x2	Avstånd till närmaste kollektivtrafikanslutning
x3	Avstånd till närmaste köpcentrum/handelsgata
x4/x5	Indikatorvariabler för kontinuerlig/sekventiell renovering
x6	Årtal för senaste renovering
x7	Antal frisörer i närområdet
x8	Antal år i samma lokal
x9	Antal år företagsnamnet funnits
x10	Antal frisörer i salongen, inklusive hyrstolar
x11	Genomsnittlig ålder på frisörerna
x12	Genomsnittlig yrkeslivserfarenhet som frisör
x13	Genomsnittligt antal kurser per år
x14	Areal för kunder i salongen

TABELL 2.1: LISTA ÖVER SAMTLIGA VARIABLER

De tre översta variablerna mättes genom att salongerna för varje kategori svarade namnet på det som ansågs närmast varpå avståndet mättes på eniro's kartor med en linjal. Ytterligare en variabel som kontrollerats med eniro är antalet frisörsalonger i närområdet där närområdet definierats som strax under 250 meters gångväg. Arealen har mätts med ögonmått av mig varför den kan upplevas som osäker, jag känner mig dock säker i min förmåga och upplever att felet borde vara lika stort vid små stora ytor.

Årtalet för senaste renovering, x6, samlades endast in för de salonger som svarade att det rådde sekventiell renovering varför den senare ströks. Av indikatorvariablerna behövdes endast en av de insamlade staplarna eftersom den andra var en linjärkombination av den första, varpå variabel x5 också ströks.

Nedan följer ett utdrag för samtliga 21 observationer, med namnen på frisörsalongerna exkluderat:

Obs	y	x1	x2	x3	x4	x7	x8	x9	x10	x11	x12	x13	x14
1	430	200	60	490	0	6	16	16	5	40	16	2	45
2	300	50	80	510	1	12	40	34	5	60	44	0	30
3	482	100	400	400	0	6	1	1	4	32	8	2	45
4	495	470	330	680	0	3	17	1	1	38	22	2	32
5	400	2	330	630	1	10	5	41	4	31	16	2	45
6	530	460	320	350	0	2	23	25	16	34	15	6	140
7	445	70	130	130	0	6	110	25	3	40	15	2	40
8	450	120	270	310	0	10	2	2	5	28	6	2	25
9	520	460	750	730	0	3	4	4	7	30	11	4	70
10	495	800	35	80	0	4	73	20	7	30	6	6	50
11	215	25	130	550	1	8	11	15	1	33	14	0	20
12	300	150	260	630	0	9	36	36	7	40	20	2	80
13	450	40	370	165	1	9	14	14	3	45	25	2	40
14	600	420	100	65	0	4	20	6	3	37	16	8	60
15	485	50	300	220	1	5	14	14	3	23	4	3	30
16	390	150	500	600	0	15	4	4	2	42	22	1	45
17	600	670	80	70	0	2	6	6	6	32	14	4	80
18	450	50	70	430	0	13	35	35	5	60	40	0	50
19	650	800	40	0	0	2	4	46	4	28	8	6	30
20	540	360	220	500	1	9	7	7	2	27	6	2	40
21	400	100	85	200	1	8	15	30	2	54	30	2	65

TABELL 2.2: DATA FÖR ANALYS.

3. METODER

3.1 REGRESSION

Vi använder oss av regression där regression definieras som att vi har en responsvariabel som linjärt kan estimeras med ett visst antal precist kända variabler samt en parameter för slump. Det vanligaste sättet att anpassa förhållandet mellan förklarande variabler och responsvariabel är genom den så kallade minsta kvadrat metoden, vilket också är den tillämpade anpassningsmetoden i detta arbete. Det skall dock för tydlighetens skull nämnas att det finns fler anpassningsmetoder med andra meriter.

3.1.1 LINJÄR REGRESSION

Linjär regression tillämpar en matematisk modell enligt följande logik (Rolf Sundberg, Tillämpad matematisk statistik, m.fl.):

$$Y = \alpha + X\beta + \varepsilon$$

Där Y är en vektor med samma längd som antalet observationer η, som innehåller de förklarade variablernas värde i varje observation. Vidare är α en η -vektor med genomsnittsvärdet för Y i varje element, X en η*ρ-matris, där ρ är antalet förklarande variabler, som innehåller alla observerade värden på de förklarade variablerna. β är en η –vektor med förhållandena sinsemellan förklarande- jämte respons-variabel.
Sist har vi en stokastisk η –vektor ε som representerar avvikelsen från det perfekta matematiska förhållandet. Varje element ε_i ska vara oberoende av övriga variabler samt sinsemellan vara oberoende fördelade enligt följande för att modellantagandet ska vara riktigt:

$$\epsilon_i \sim N(0,\sigma^2) \quad i \in [1,\eta]$$

För något okänt σ.

3.2 MINSTA KVADRAT ANPASSNING

Under minsta kvadrat metoden observerar vi, efter att ha valt värden för alla förhållanden β, vilket utfall e_i vi får som värde från de stokastiska variablerna ϵ_i genom följande:

$$e_i = y_i - \hat{y}_i$$

Där y_i och $\hat{y}_i$ står för observerat värde för observation i inuti vektorn Y respektive det värde som modellen föreslår för samma observation i som ett resultat av de valda parametrarna β. Målet är givetvis att de värden den matematiska modellen förslår inte ska ligga alltför långt ifrån de verkliga observerade värdena, problematiken uppstår då vi inte kan välja en modell som passar perfekt för samtliga värden.

I och med att de observerade värdena e_i ändras beroende på vilka förhållanden vår modell väljer uppstår ett krav på en bedömningsprocess som väljer vilken anpassning som var bäst utifrån det processvalet. Minsta kvadrat metoden formulerar en sådan bedömning genom att observera den kvadrerade summan av alla residualer e_i, det vill säga:

$$RSS = \sum_{\eta} e_i^2$$

Modellen anses anpassad när β välj så att RSS minimeras. Rent praktiskt medför kvadreringen av residualerna att vi erhåller en modell som föredrar flera små avvikelser gentemot en större, även om den sammanlagda längden på avvikelserna skulle vara densamma.

3.3. KORSVALIDERING

Korsvalidering är en metod som lämpar sig väl då man inte bara försöker finna samband i form av värden för x-variabler, utan även känner att det är önskvärt med ett mått på hur väl sambanden lämpar sig för att prediktera ett värde utanför de insamlade datapunkterna. I praktiken innebär det att man utelämnar en given del av den insamlade datan för att endast skatta sin modell med hjälp av de resterande observationerna. Det utlämnade datasetet blir då ett valideringsset och metoden upprepas med olika grupperingar så att alla variabler används i

valideringsetet en gång. Genom denna metod får man ett mått på hur väl modellen predikterar nya datapunkter genom att summera kvadraterna på varje felavstånd i och för varje gruppering. Denna summa kallas Predicted Residual Sum of Squares eller PRESS:

$$PRESS = \sum_{m} \sum_{i \subseteq I_m} e_{(i)}^2$$

Där $e'_{(i)}$ är avvikelsen mellan observationen och det värde modellen förutsäger, m antalet olika grupperingar och I_m är gruppen av observationer som inte avändes för att skatta modellen under gruppering m.

När man jämför modeller med hjälp av PRESS kan man gott låta summan vara obehandlad och observera hur den rör sig då man inkluderar eller exkluderar variabler, dock finns det ett intresse av att kunna jämföra detta tal med standardavvikelsen för olika modeller varför vi också kan vara intresserade av det typiska prediktionsfelet i den enhet modellen avser. Det typiska prediktionsfelet fås, i likhet med standardavvikelse, genom:

$$Typiskt\ prediktionsfel = \sqrt{PRESS/n}$$

Under modelleringen av data kommer vi även titta på modeller där vi logaritmerar responsvariabeln y, varpå det typiska prediktionsfelet istället erhålles som det typiska förhållandet mellan predikterat värde och verkligt värde enligt:

$$\frac{y}{\hat{y}} \approx e^{\sqrt{PRESS/n}}$$

3.3.1 LEAVE ONE OUT KORSVALIDERING

Leave One Out är ett specialfall av korsvalidering där man tillåter storleken på valideringsgruppen att bara innehålla en observation, vilket resulterar i att man får lika många

grupper som man har observationer eftersom varje observation utgör valideringsgrupp precis en gång. Vårt matematiska uttryck förkortas då något till följande:

$$PRESS = \sum_{i} e_{(i)}^{2}$$

Där i avser både grupp samt utelämnad observation. Metoden appliceras enligt samma procedur även när responsvariablerna är logaritmerade.

3.4 ALGORITMER, SAS

Vi kommer att i huvudsak använda oss av en algoritm i SAS för att få fram alternativ till modeller men då de andra även tillämpas och framförallt står till grund för den sista algoritmen förklaras samtliga.

3.4.1 FORWARD SELECTION

Forward Selection sker i SAS genom att programvaran startar med responsvariabeln allena, för att sedan göra en regression på varje förklarande variabel och välja ut den som var mest statistiskt signifikant. Efter att ha funnit den variabeln så undersöker algoritmen vilken av de ytterligare variablerna som är mest statistiskt signifikant för att sedan inkludera den. Metoden upprepas tills programvaran noterar att den bästa bland resterande signifikanser inte överstiger en användar angiven gräns.

3.4.2 BACKWARD ELIMINATION

Backward Elimination är i någon mening motsatsen till föregående. Här börjar vi med att göra en analys på hela underlaget med samtliga av de förklarande variablerna inkluderade, för att sedan exkludera den variabeln som har lägst signifikans. Modellen estimeras efter det om och ytterligare en variabel exkluderas efter samma krav. Metoden stannar då signifikansen för nästa variabel att exkluderas överstiger en användarangiven gräns.

3.4.3 STEPWISE REGRESSION

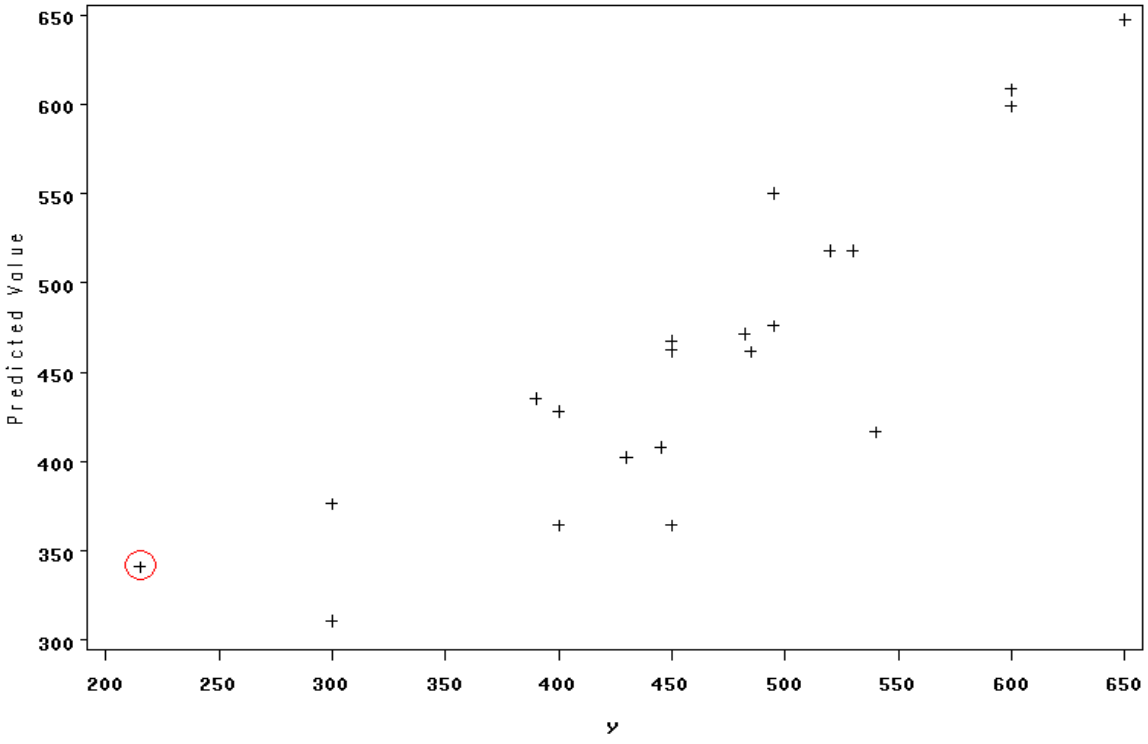
Stepwise regression är, som utlovat, en kombination av de två föregående algoritmerna. Metoden börjar med endast responsvariablerna och utför en inkludering enligt samma metod

som Forward Selection, för att därefter genast försöka göra en Backward Elimination på den nya gruppen variabler. Om någon av metoderna inte kan genomföras då ingen variabel möter någon av gränserna fortsätter proceduren genom att applicera samma algoritm ytterligare en gång. Algoritmen stannar antingen när ingen av de två ovanstående algoritmerna kan utföra ytterligare en exkludering/inkludering eller när algoritmen exkluderar och inkluderar samma variabel i oändlighet.

4. PLOTTNING OCH EXKLUDERINGAR

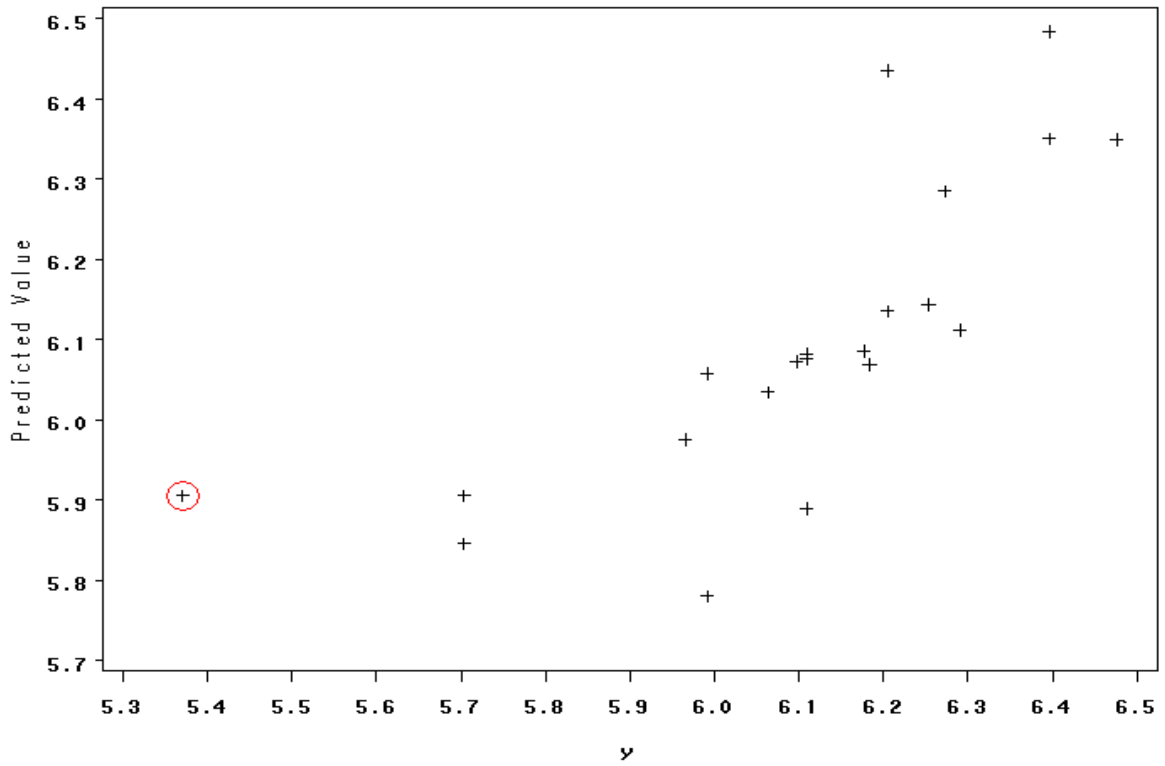
Bland de första noteringar som går att finna bara genom att titta på dataunderlaget är att observationen med det lägsta värdet på responsvariabeln pris ligger väldigt långt från övriga observationer. Vi får undersöka om den prisklassen skiljer sig så pass mycket att den av någon anledning inte kan skattas med samma parametrar som de övriga observationerna.

Om vi betraktar en plot som visar en regression över alla variabler ser vi följande:



FIGUR 4.1: REGRESSIONSPLOT, SAMTLIGA VARIABLER INKLUDERADE

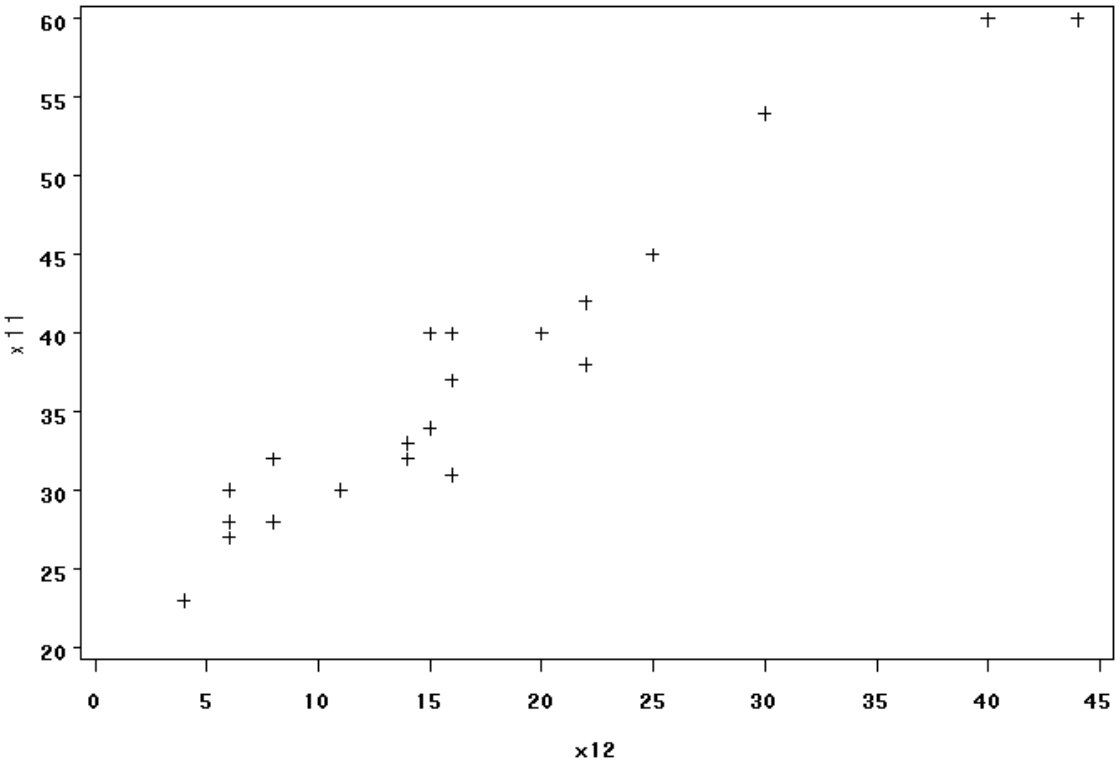
Ovanstående effekt blir tydligare när vi arbetar med att reducera antalet variabler för modellen, samt även när vi arbetar med en logaritmerad responsvariabel, se nedan:



FIGUR 4.2: REGRESSIONSPLOT, LOGARITMERAT SLUTPRIS, SAMTLIGA VARIABLER INKLUDERADE

Med ovanstående som grund kommer den observationen strykas från all modellering. Det är viktigt att poängtera att vi då krymper intervallet för vilket vi kan anta att vi kan göra statistiskt korrekta predikteringar för priset.

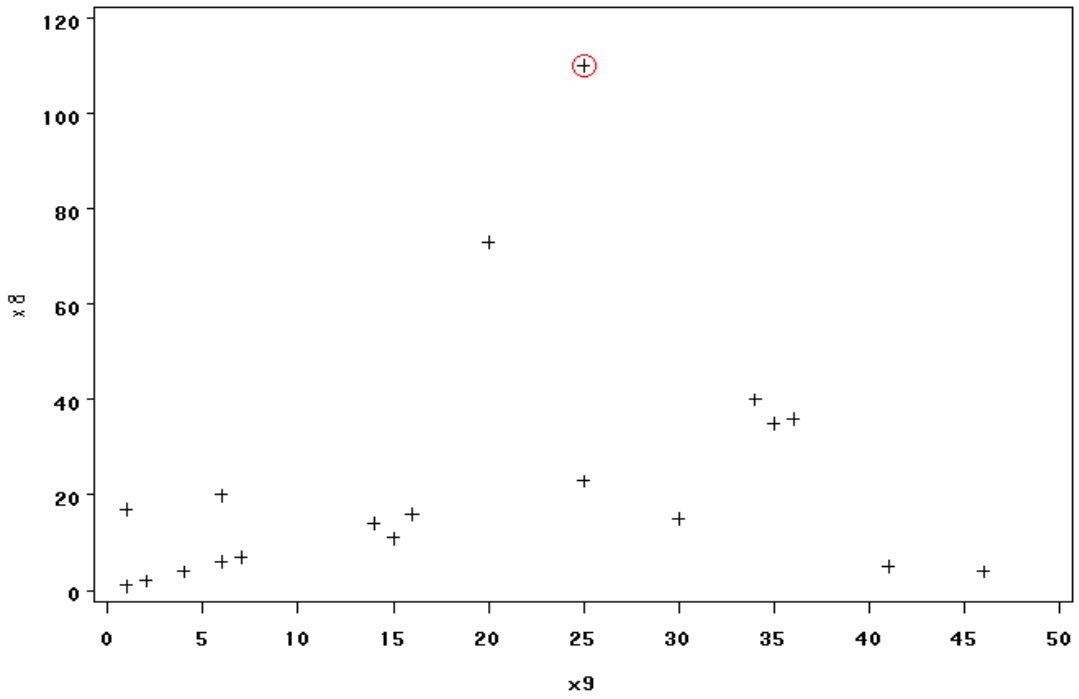
Ytterligare en effekt bland data som märks ganska snabbt vid plottning, samt eventuellt med blotta ögat, är att en del variabler samvarierar. Betrakta till exempel x11 ihop med x12, där de variablerna stod för snittålder respektive snitterfarenheten bland frisörerna på salongen. Av naturliga själ samvarierar de variablerna:



FIGUR 4.3: X11 PLOTTAT MOT X12

Det är uppenbart att det föreligger ett mycket starkt samband och det kan därför bli svårt för SAS eller annan mjukvara att bedöma vilken variabel det egentligen är som står för förklarandet av variationen i data. För detta variabel-par har jag valt att i analysen alltid undersöka resultatet av att växla variabel om SAS valt den ena av dem.

Vidare finner vi under plottning att x8 samt x9 som stod för tid i lokalen respektive hur länge namnet funnits uppvisar samvariation:



FIGUR 4.4: X8 PLOTTAT MOT X9

Här ser vi ett tydligt, om än inte lika tydligt, samband mellan de två variablerna. Vidare kan vi lägga märke till ett extremvärde på 110 år för tid i samma lokal. Ett sådant extremvärde kan ha olyckliga konsekvenser när man försöker anpassa ett värde på förhållandet mellan slutpris och samma variabel, varför jag i det här fallet valt att exkludera variabel x8 från modelleringen då vi med ganska stor säkerhet kan förlita oss på att en stor del av samma information finns fångad i variabeln x9 som är fri från sådana extremvärden.

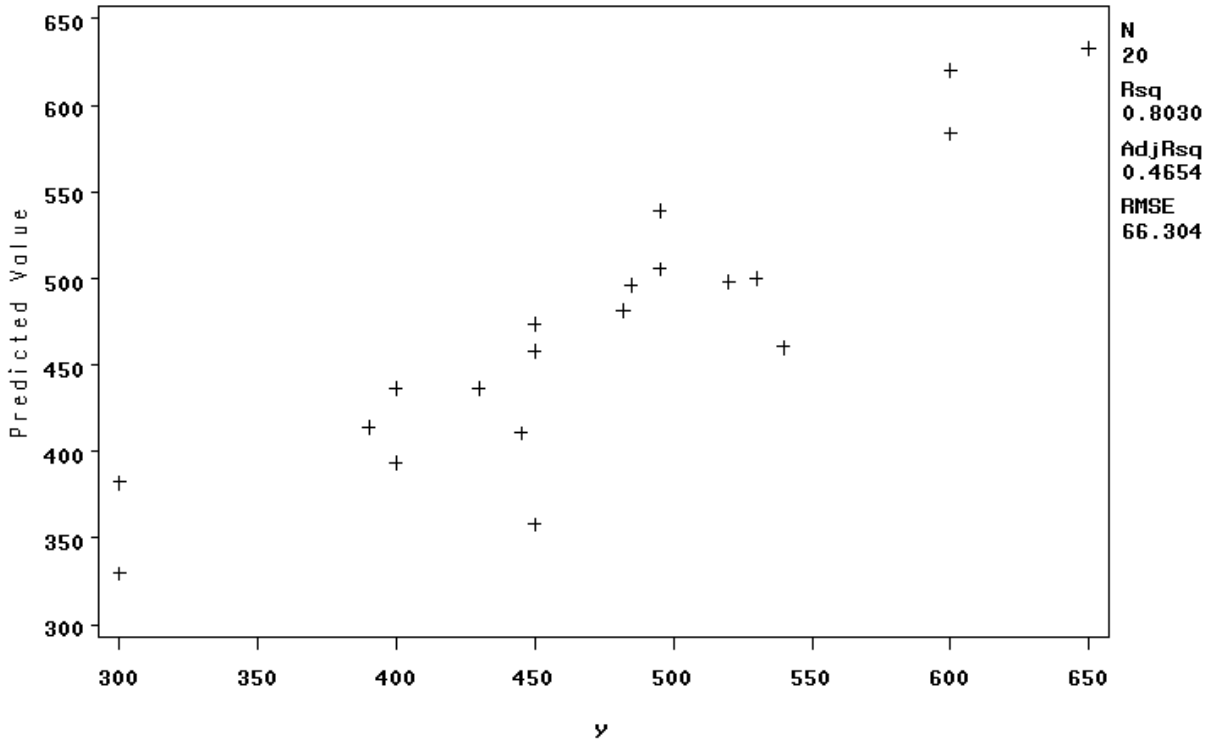
I övrigt finner inte jag några andra frågetecken att ta ställning till innan vi kan börja försöka anpassa en eller flera modeller på underlaget.

5. STATISTISK MODELLERING

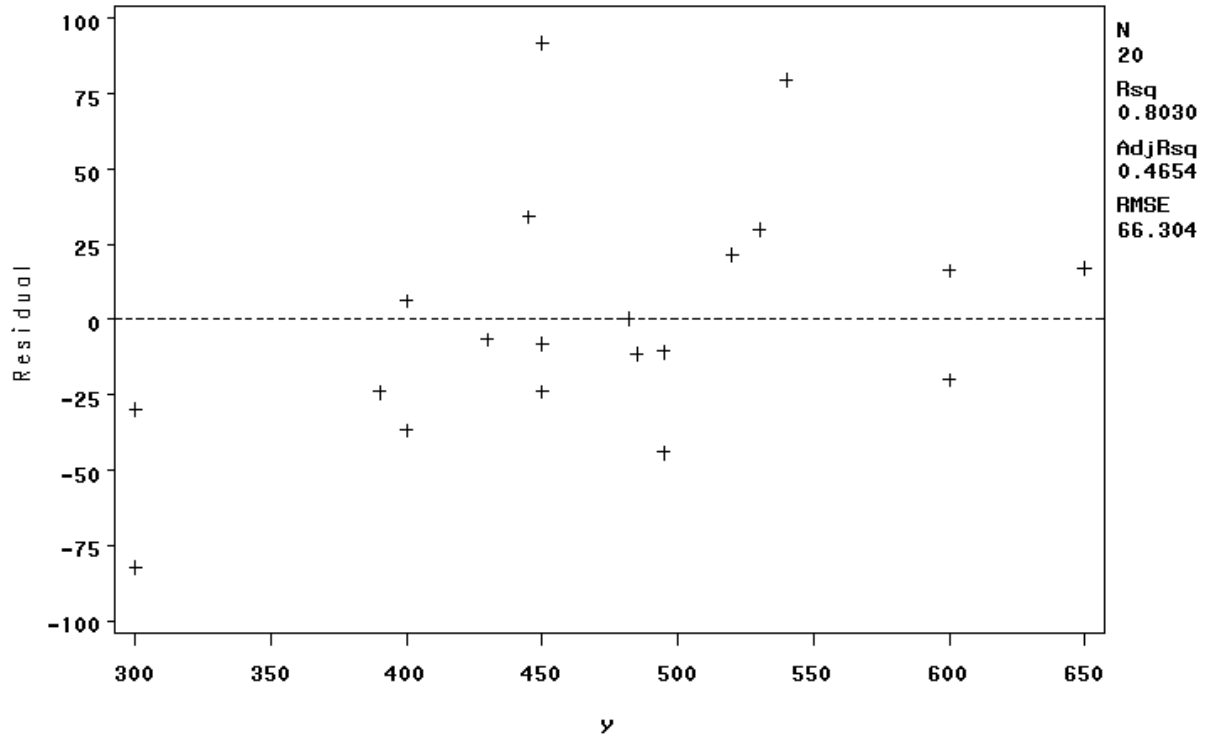
5.1 MODELLERING: LINJÄR REGRESSION

I denna sektion kommer vi modellera underlaget utan att göra några transformeringar på vare sig respons- eller förklarande-variabler. Vi kommer låta programvaran SAS succesivt inkludera alternativt exkludera variabler med olika krav på signifikans för att på så sätt försöka urskilja vilka variabler som verkligen förklarar variationen i data. Vi kommer ändå presentera fler än en modell då vi senare ska göra korsvalidering på modellerna.

Vi börjar med att presentera två plottar samt en tabell med information för en regression gjord över samtliga variabler:



FIGUR 5.1: PREDIKTERAT VÄRDE MOT VERKLIGT, ALLA FÖRKLARANDE VARIABLER



FIGUR 5.2: RESIDUALER, ALLA FÖRKLARANDE VARIABLER

Modell	Valda variabler	Förklaringsgrad R^2	Standardavvikelse σ
Total	Alla	0,80	66,3

TABELL 5.1.1: MODELL-INFORMATION, SAMTLIGA VARIABLER INKLUDERADE

Bland graferna kan vi i synnerhet notera de två observationerna med lägst verkligt pris som denna modell då estimerar till ett högre pris givet de förklarande variablerna. Denna effekt för de två observationerna kommer hålla i sig i samtliga modeller, även där vi transformerar responsvariabeln genom logaritmering, vilket möjligen föreslår att modellen inte sträcker sig ner till den prisklassen heller. Jag har ändå valt att inkludera punkterna även om det kan vara så att någon bit av information gällande dom inte fångats i underlaget då bristen på observationer medför att de står för 10% av det resterande underlaget.

Gällande tabellen så kan vi notera en relativt hög förklaringsgrad på 0,80.

Inför modelleringen där vi nu ska reducera variabelurvalet så skall det nämnas att modellering även har skett utan hjälp av algoritmer men då resultaten blir snarlika och det inte föreligger någon intuitiv skillnad mellan de resultaten och de valda av algoritmer har jag valt att använda mig av algoritmer som verktyg för att välja modeller. Då olika gränser tillåter olika många variabler och vi är intresserade av just olika modeller att senare korsvalidera presenteras följande tabell över vilka gränser som använts samt en överblick över de resulterande modellerna. Fler undersökningar har gjorts med alla tre algoritmer som beskrivits, men då de ofta ger samma resultat visas bara en av utsökningarna som kan användas för att få fram samma modell.

Modell	Signifikans, gränser		Valda variabler	Förklaringsgrad R^2	Standardavvikelse σ
	Inkludering	Exkludering			
Modell A	0,20	0,05	x13	0,53	63,91
Modell B	0,10	0,10	x1 x13	0,61	60,13
Modell C	0,15	0,20	x1 x3 x9	0,67	56,64
Modell D	-	0,10	x1 x2 x3	0,67	56,41
Modell E	0,25	0,25	x1 x3 x9 x11	0,71	54,84
Modell F	0,25	0,25	x1 x3 x7 x9	0,71	55,23

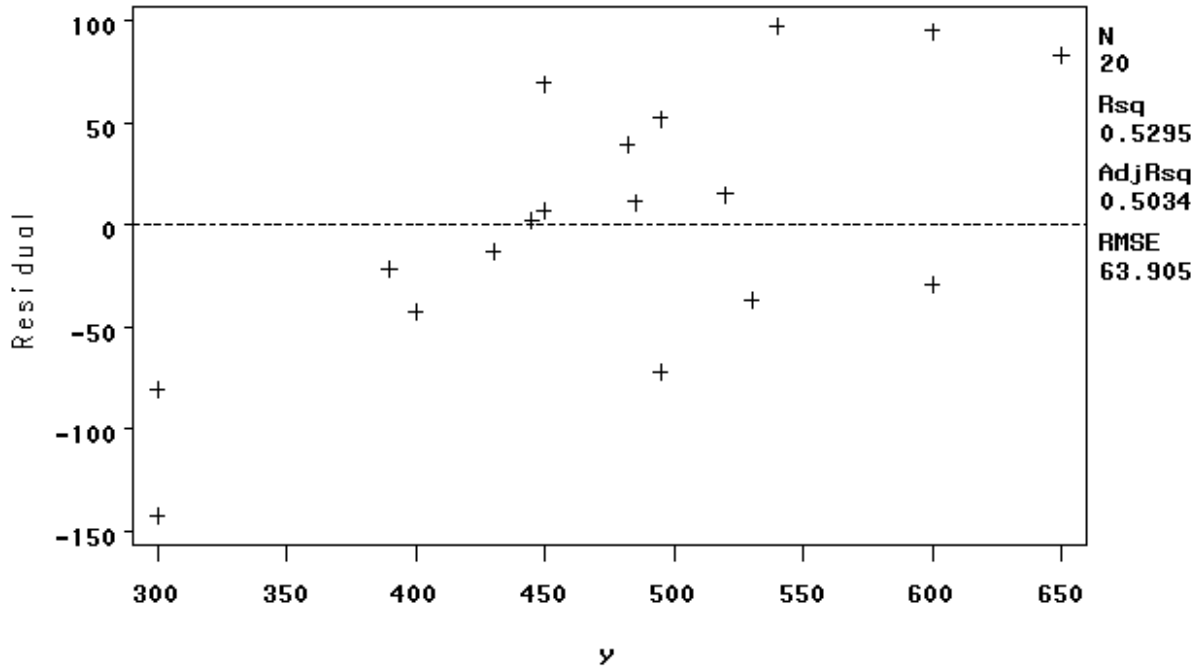
TABELL 5.1.2: MODELLER SOM ETT RESULTAT AV GRÄNSER FÖR ALGORITMER

Ovan kan vi se att variabeln x13 väljs om algoritmen inte får gå fler steg, för att sedan exkluderas om fler iterationer fortlöper. Vi kan notera i Modell E samt F att gränserna för signifikans är något högre än önskvärt men lägre gränser än så för stepwise-algoritmen resulterade endast i modeller identiska med Modell C. Att samma gränser använts för modell F med olika resultat förklaras genom att x11 alternerades med x12 enligt förklaring från inledande plottning. Vidare kan vi notera att en stor del av variationen i underlaget blir förklarad trots det kraftigt reducerade antalet förklarade variabler, notera i synnerhet den lilla förlusten av att exkludera x11 från Modell E och därmed hamna i Modell C. Intressant är att samtliga modeller har lägre standardavvikelse än den modell där alla förklarande variabler är inkluderade, i synnerhet alla modeller efter och inklusive modell C. Nedan presenteras detaljerad information för de inkluderade variablerna i respektive modell:

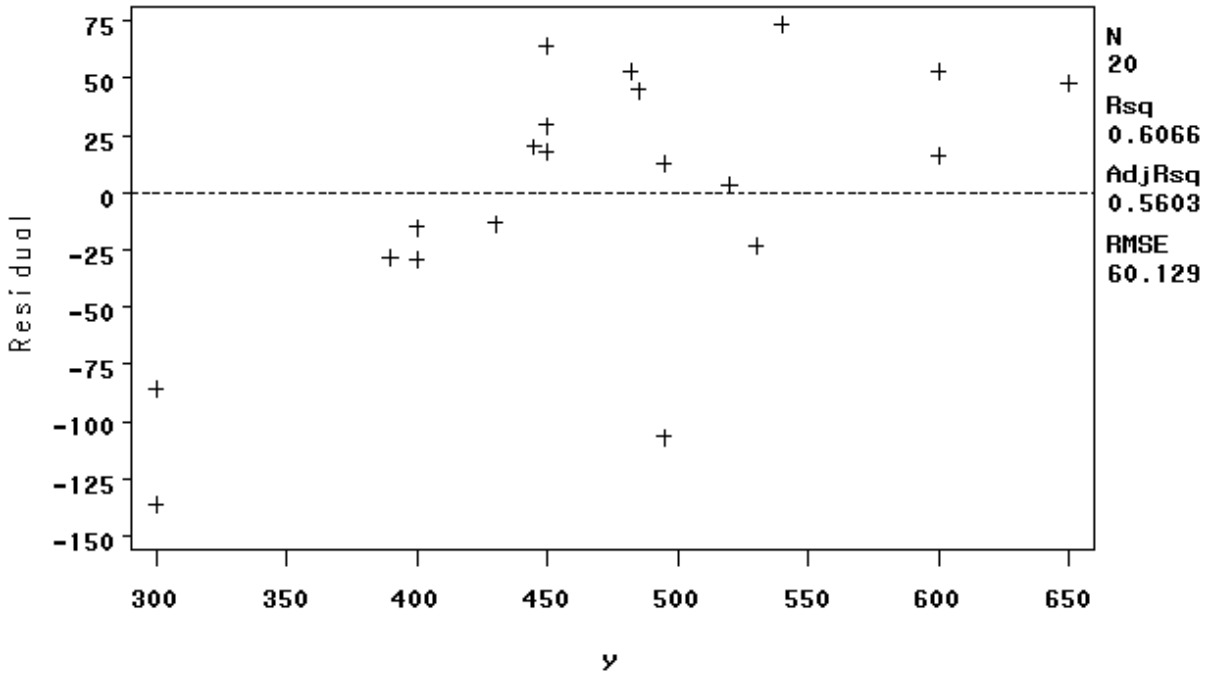
	Variabler			
Modell	Kod	Text	Koefficient	Signifikans
Modell A	x13	Kurser per år	31,05	0,0003
	x1	Avstånd till Konkurrent	0,14	0,0856
Modell B	x13	Kurser per år	18,19	0,0766
	x1	Avstånd till Konkurrent	0,20	0,0017
	x3	Avstånd till köpcentrum	-0,14	0,0391
Modell C	x9	År namnet funnits	-1,59	0,0956
	x1	Avstånd till Konkurrent	0,21	0,0010
	x2	Avstånd till tunnelbana	0,16	0,0885
Modell D	x3	Avstånd till köpcentrum	-0,20	0,0160
	x1	Avstånd till Konkurrent	0,17	0,0093
	x3	Avstånd till köpcentrum	-0,13	0,0379
	x9	År namnet funnits	-1,12	0,2451
Modell E	x11	Snittålder	-2,04	0,1713
	x1	Avstånd till Konkurrent	0,14	0,0596
	x3	Avstånd till köpcentrum	-0,11	0,0744
	x7	Konkurrenter i närområdet	-4,79	0,1995
Modell F	x9	År namnet funnits	-0,89	0,1383

TABELL 5.1.3: SIGNIFIKANSER FÖR DE INKLUDERADE VARIABLERNA

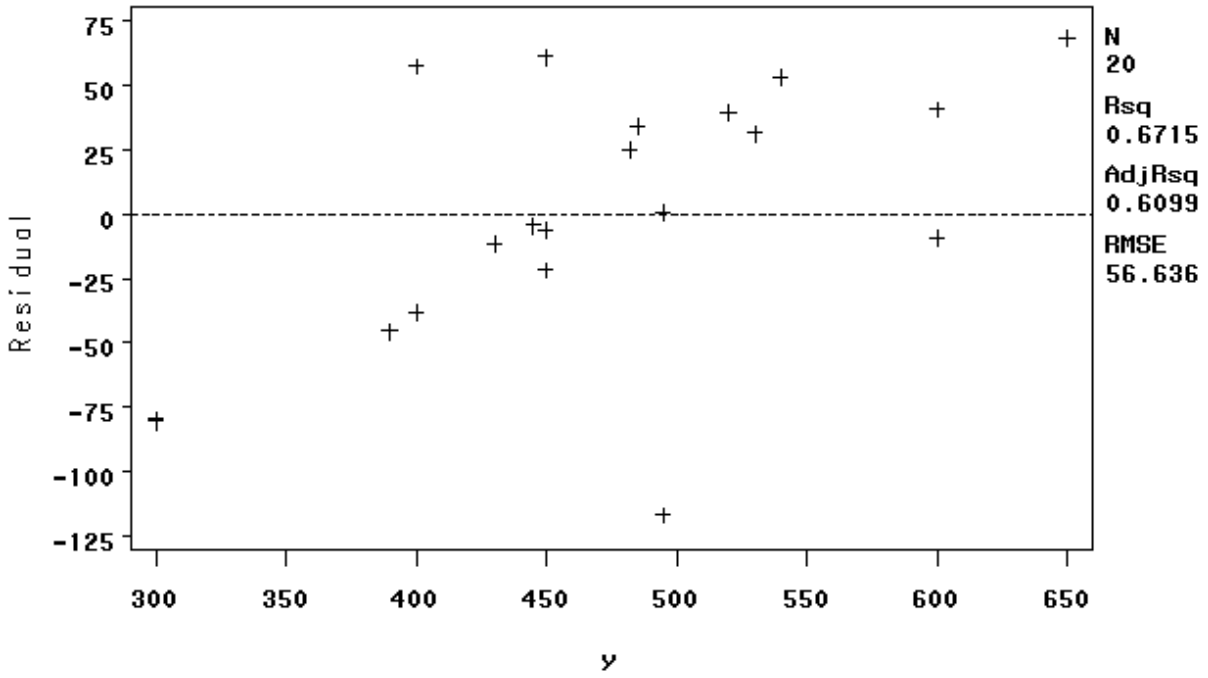
Koefficienterna för x1, x3 samt x9 verkar relativt stabila. Samtliga modeller lider dock av residualer som inte ser alltför normalfördelade ut, vilket är ett viktigt antagande för linjär regression. Detta avhjälps något när vi övergår till logaritmerade responsvariabler. Nedan presenteras residualplottar för samtliga av de ovanstående modellerna.



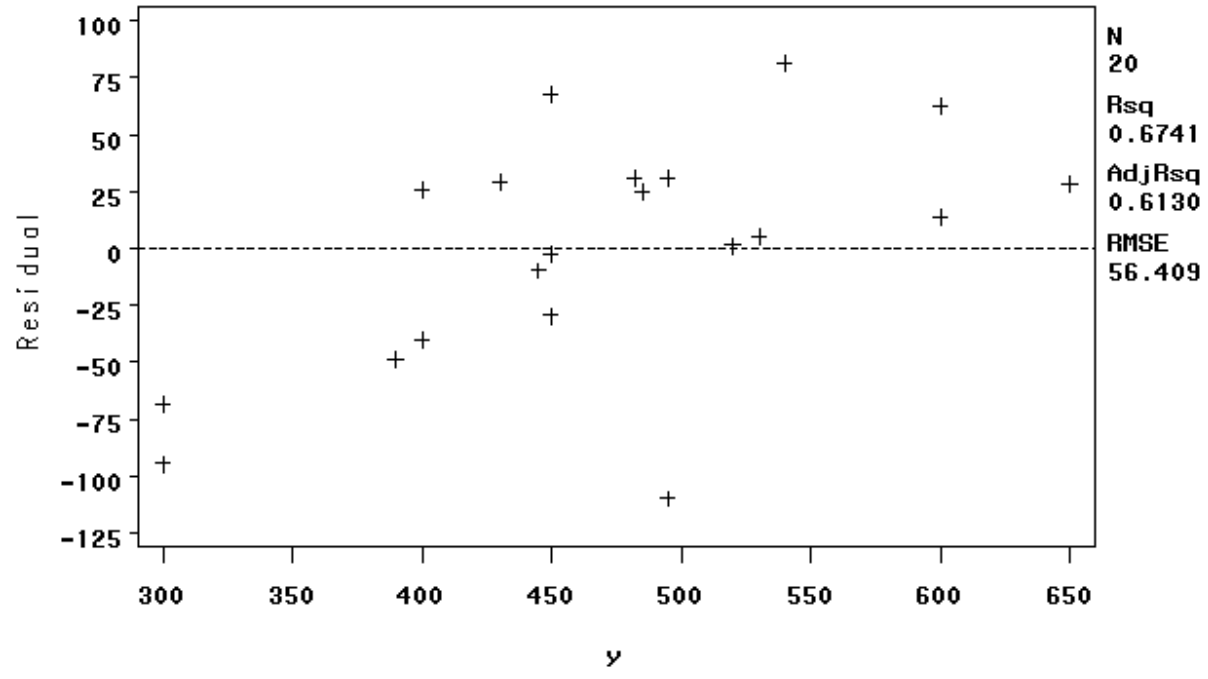
FIGUR 5.3: RESIDUALER, MODELL A



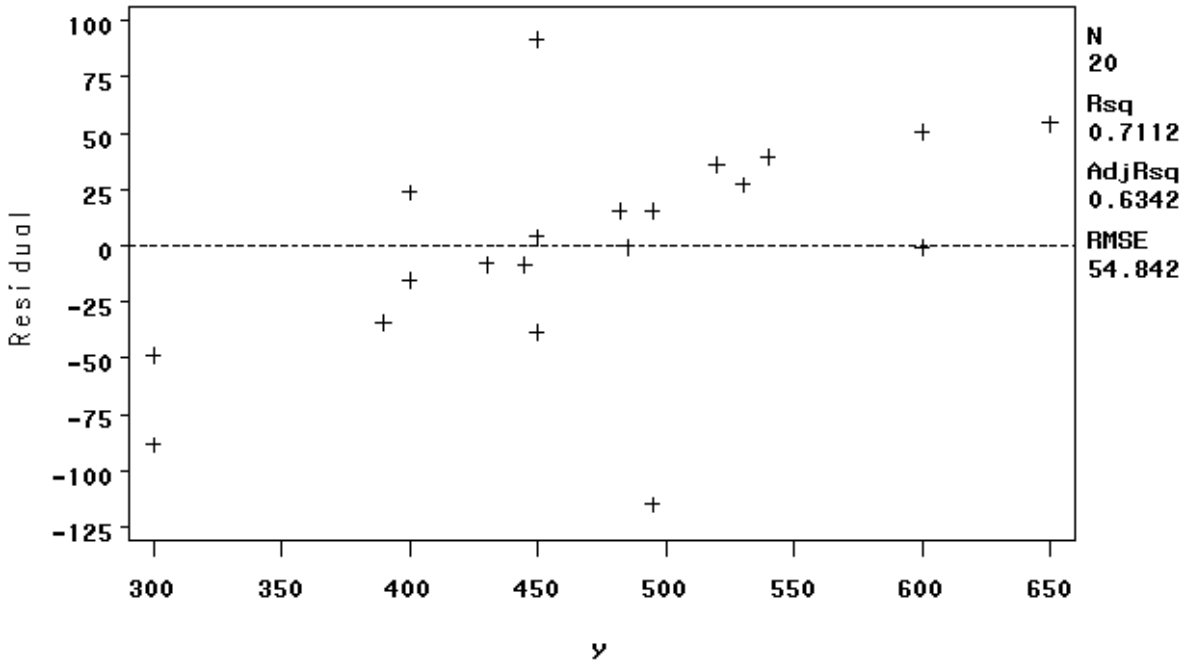
FIGUR 5.3: RESIDUALER, MODELL B



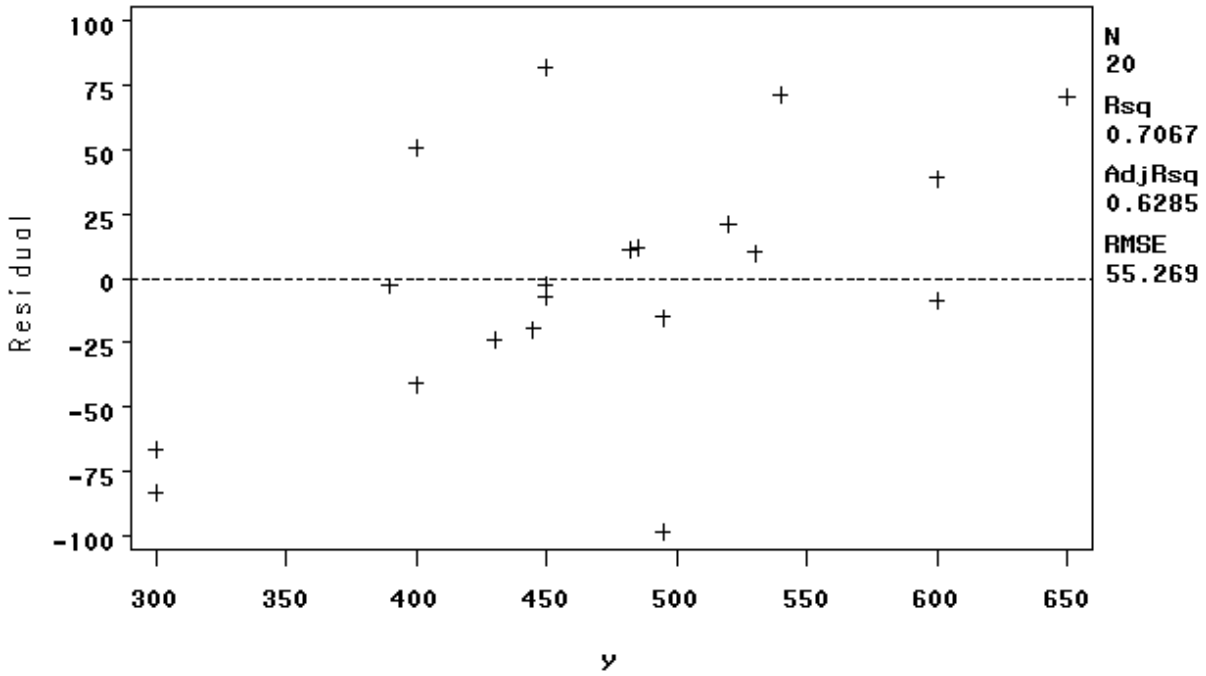
FIGUR 5.3: RESIDUALER, MODELL C



FIGUR 5.3: RESIDUALER, MODELL D



FIGUR 5.3: RESIDUALER, MODELL E



FIGUR 5.3: RESIDUALER, MODELL F

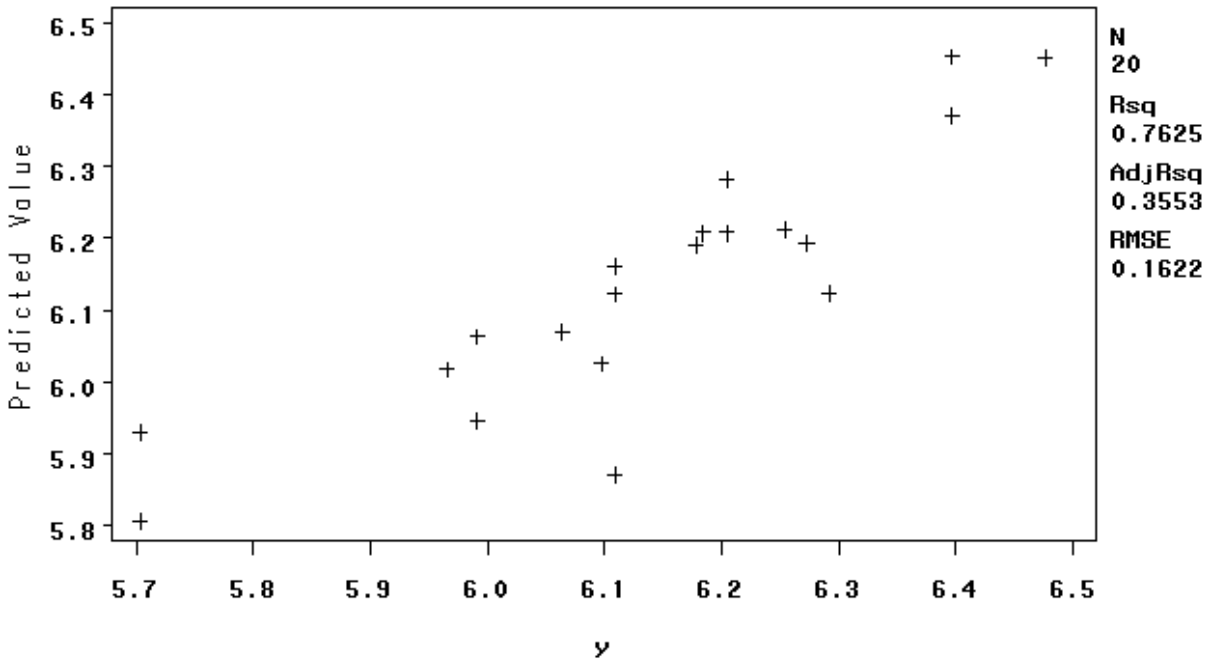
Residualerna för Modell A samt F ser mest ut att kunna passera som normalfördelade data. Framförallt är det de två första observationerna, som tidigare nämnts, vars pris alltid blir överskattat.

5.2 MODELLERING: LINJÄR REGRESSION, LOGARITMERAD RESPONSVARIABEL

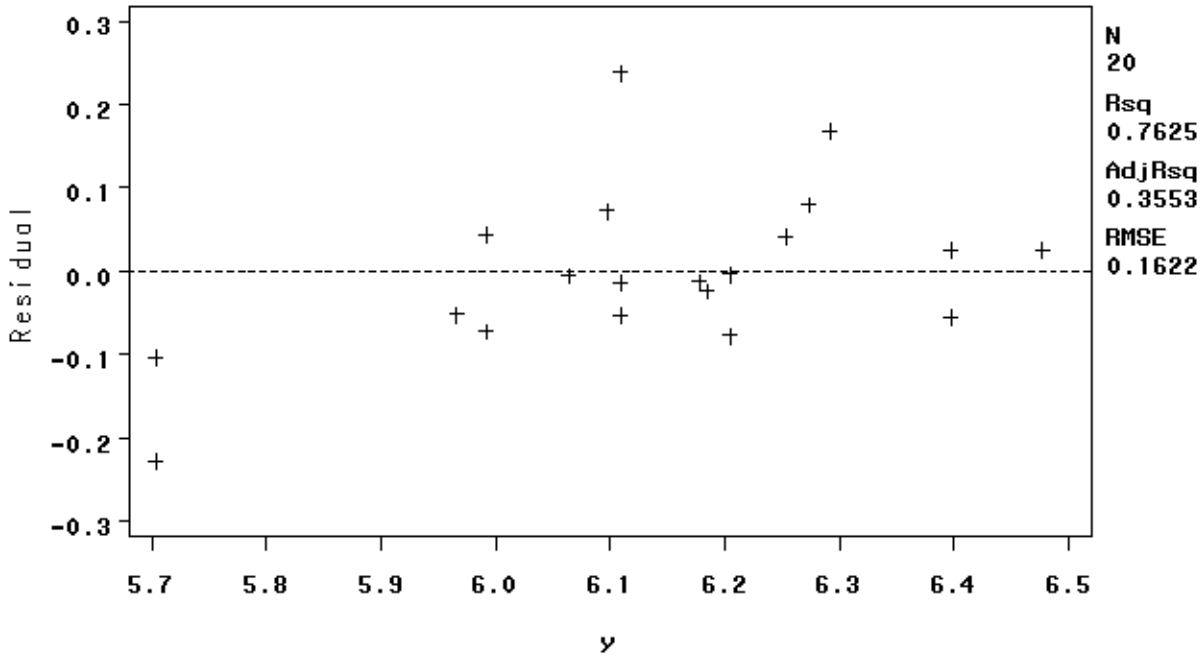
I följande sektion har vi logaritmerat responsvariabeln, varpå resultaten upphör att vara jämnförbara med den föregående analysen. Vi väljer att logaritmera responsvariabeln då en linjär modell kan ifrågasättas när vi jobbar med pengar som responsvariabel. Modeller med logaritmerad responsvariabel fångar, istället för absoluta värden, proportionella ändringar för responsvariabeln som ett resultat av ändrade förutsättningar bland de förklarande variablerna. Betänk till exempel möjligheten att en ytterligare avklarad frisörkurs inte alltid skulle resultera absolut påslag på priset, utan istället ett proportionellt. Resonemang av det slaget fångas i dessa modeller.

I övrigt har data behandlats på samma sätt som tidigare genom applicering av de olika regressions-algoritmerna med olika gränser för exkludering respektive inkludering. Modellnamnen fortsätter för tydlighetens skull från och med modell G.

För att på enklaste sätt vissa hur detta underlag ter sig gentemot regression presenteras återigen först plottar samt information för en regression på hela underlaget med samtliga förklarande variabler:



FIGUR 5.2.1 REGRESSIONSPLOT, LOGARITMERAT SLUTPRIS, SAMTLIGA VARIABLER INKLUDERADE



FIGUR 5.2: RESIDUALER, LOGARITMERAT SLUTPRIS, SAMTLIGA VARIABLER INKLUDERADE

Modell	Valda variabler	Förklaringsgrad R^2	Standardavvikelse σ
Total	Alla	0,76	0,1622

TABELL 5.2.1: MODELL-INFORMATION, LOGARITMERAT SLUTPRIS, SAMTLIGA VARIABLER INKLUDERADE

I linje med tillvägagångssättet för den icke logaritmerade responsvariabeln undersöker vi nu möjligheter att reducera antalet parametrar med hjälp av samma algoritmer som tidigare.

Modell	Signifikans, gränser Inkludering	Exkludering	Valda variabler	Förklaringsgrad R^2	Standardavvikelse σ
Modell G	0,20	0,05	x7	0,48	0,1490
Modell H	0,25	0,15	x1 x7	0,55	0,1427
Modell I	-	0,15	x1 x3 x9	0,64	0,1321
Modell J	0,20	-	x1 x3 x7 x9	0,68	0,1286

TABELL 5.2.2: MODELLER SOM ETT RESULTAT AV GRÄNSER FÖR ALGORITMER

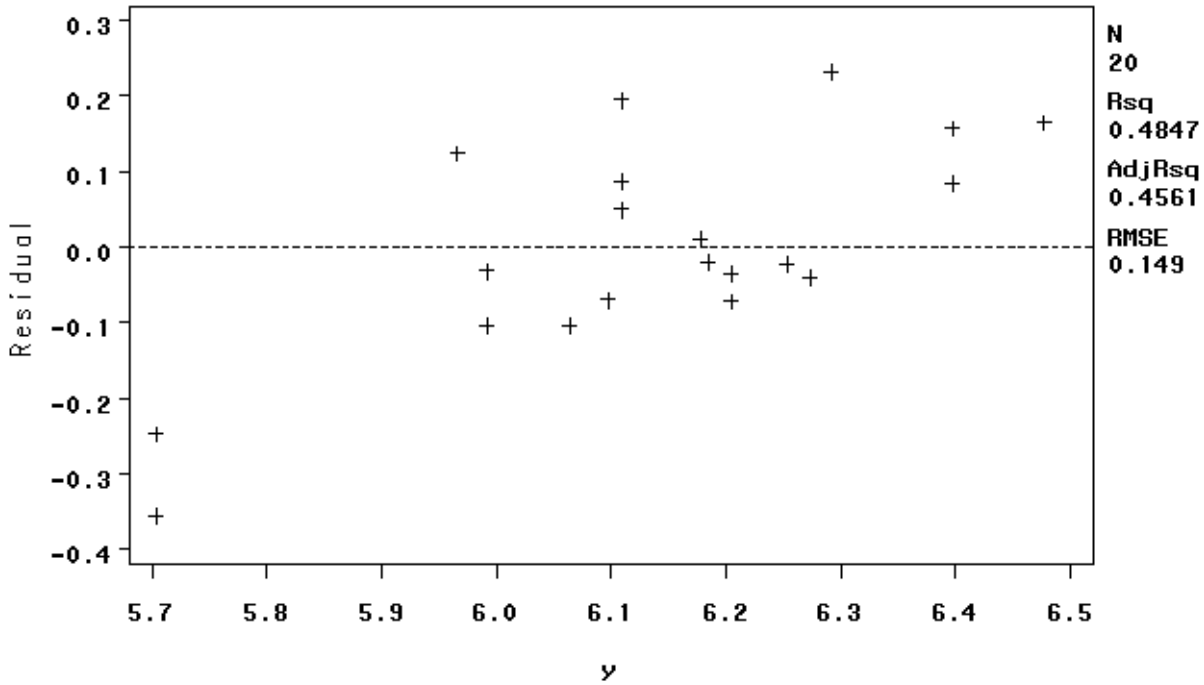
Vi ser genast att vi får färre modeller än förra gången, vilket beror på att nästan alla sökningar där man börjar öka gränserna för inkludering respektive exkludering resulterar i Modell I eller Modell J, tills man höjt gränserna så pass mycket att de förlorar sin mening. Vidare ser vi återigen att alla framtagna modeller har en lägre standardavvikelse än en modell över hela underlaget, då i synnerhet efter och inklusive Modell I. Läsare med gott minne lägger även märke till att urvalet av dessa variabler som förklarande förekom även i analysen av den icke-logaritmerade responsvariabeln y.

För att lättare bedöma modellernas riktighet presenteras återigen en mer detaljerad tabell över de inkluderade variablernas signifikans samt textförklaring.

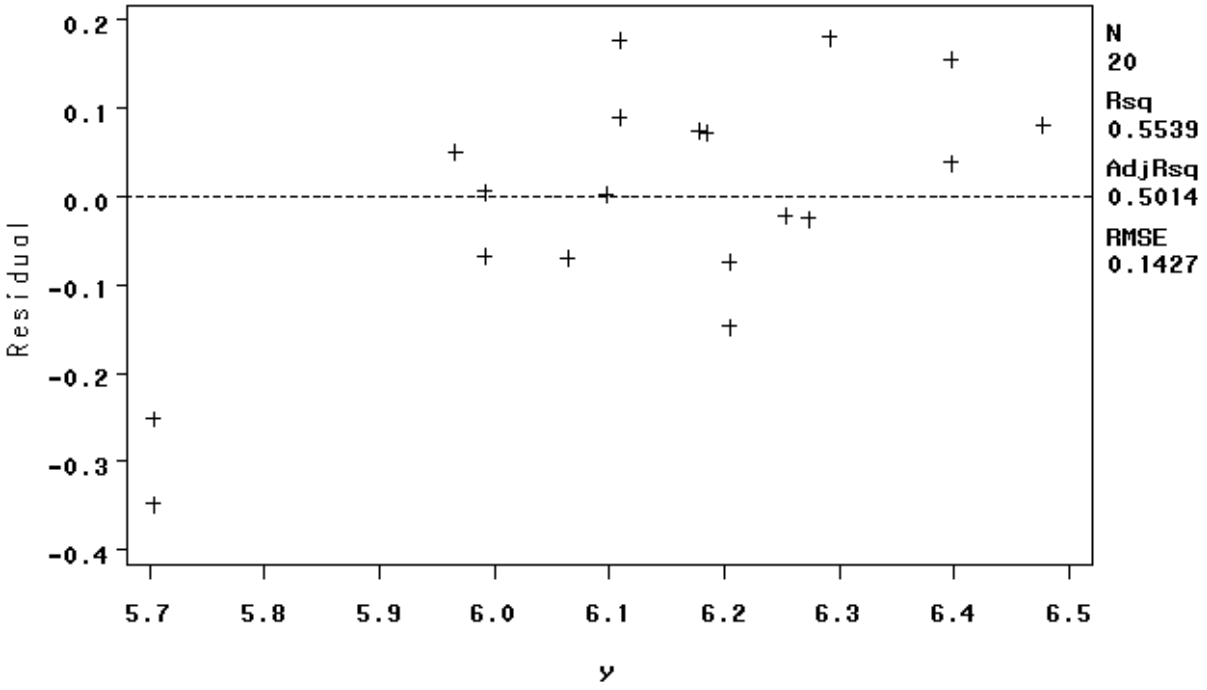
Modell	Variabler Kod	Text	Koefficient	Signifikans
Modell G	x7	Konkurrenter i närområdet	-0,03615	0,0007
	x1	Avstånd till Konkurrent	0,000289	0,1229
Modell H	x7	Konkurrenter i närområdet	-0,02245	0,0766
	x1	Avstånd till Konkurrent	0,000410	0,0048
	x3	Avstånd till köpcentrum	-0,000308	0,0437
Modell I	x9	År namnet funnits	-0,00437	0,0536
	x1	Avstånd till Konkurrent	0,000264	0,1226
	x3	Avstånd till köpcentrum	-0,000262	0,0826
	x7	Konkurrenter i närområdet	-0,01533	0,1891
Modell J	x9	År namnet funnits	-0,00389	0,0795

TABELL 5.2.3: SIGNIFIKANSER FÖR DE INKLUDERADE VARIABLERNA

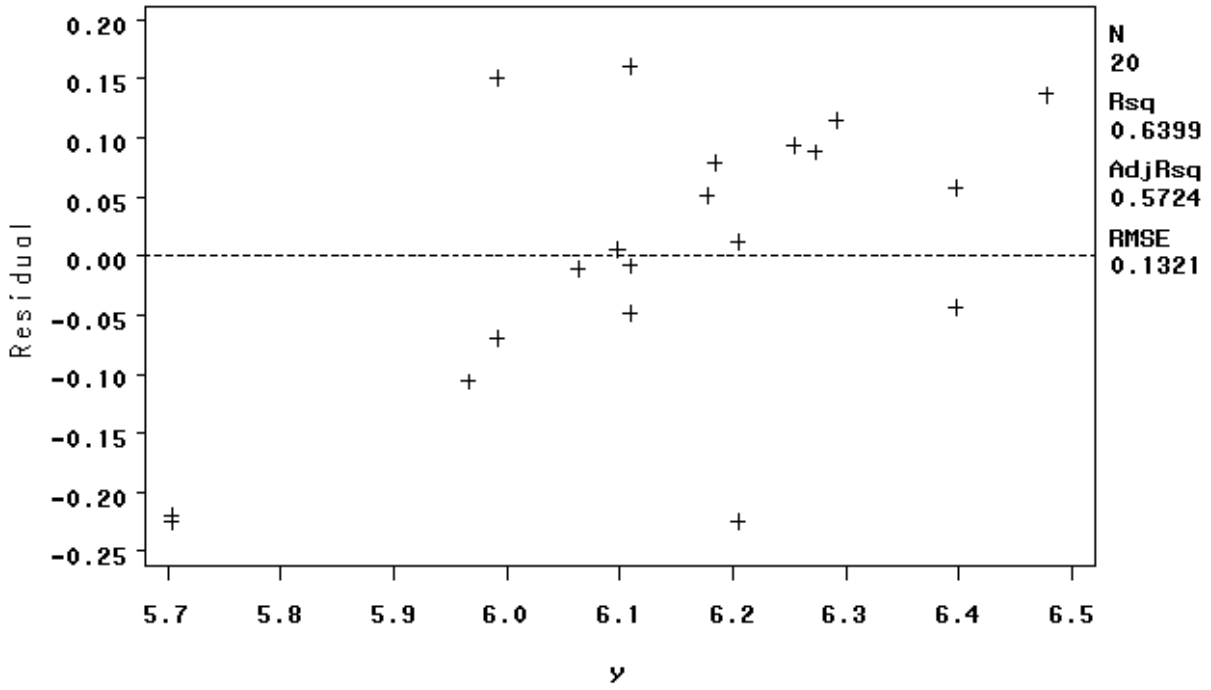
Rent signifikansmässigt ser Modell I mest attraktiv ut bland de ovan angivna modellerna. Detta kombinerat med en relativt hög förklaringsgrad gör att det blir en god kandidat för bäst lämpade modell. Vi måste dock se hur bra den och övriga modeller är på att prediktera nya värden med hjälp av korsvalidering i kommande delar av analysen. Nedan följer residualplottar för samtliga modeller då vi behöver kontrollera vilka som mest beter sig enligt ett normalfördelningsantagande bland residualerna.



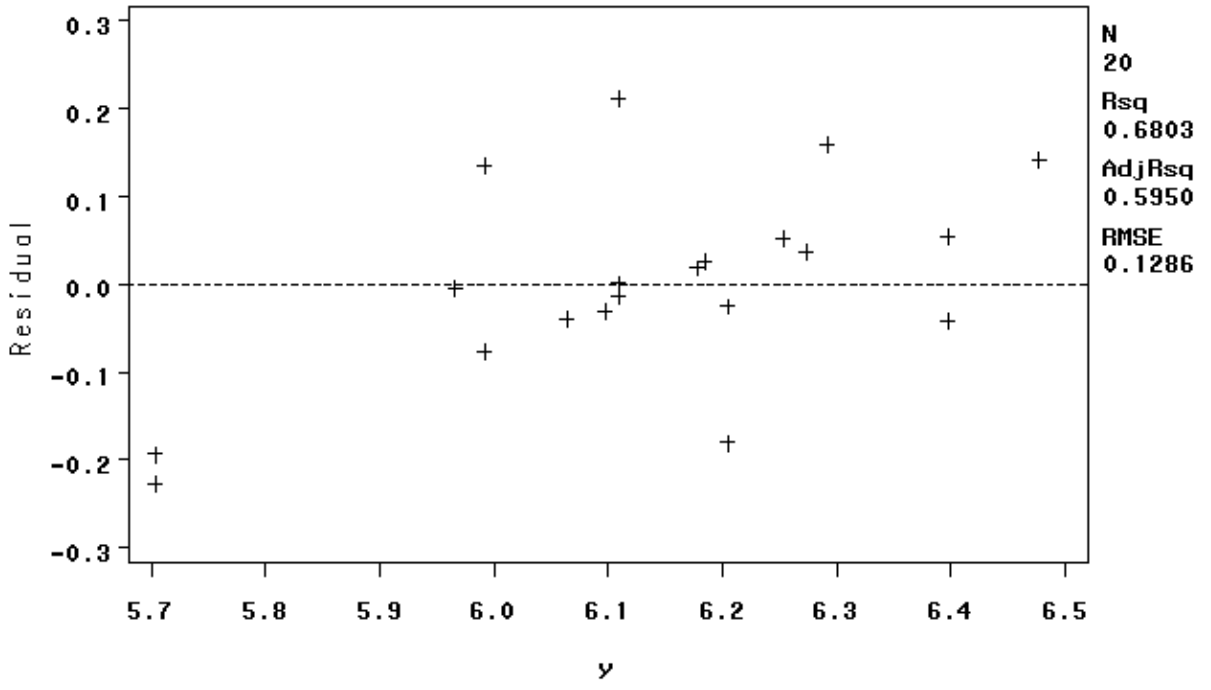
FIGUR 5.3: RESIDUALER, MODELL G



FIGUR 5.3: RESIDUALER, MODELL H



FIGUR 5.3: RESIDUALER, MODELL I



FIGUR 5.3: RESIDUALER, MODELL J

Bland plottarna ovan ser residualerna för Modell J mest slumpmässigt fördelade ut.

6. KORSVALIDERING

I denna del undersöker vi hur PRESS-värden för de olika modellerna ter sig när vi låter alla valda modeller utsättas för korsvalidering med leave-one-out metoden. Eftersom PRESS-värden för modeller med en logaritmerad responsvariabel inte är jämförbara med icke-logaritmerade kommer resultaten gruppvis.

6.1 PRESS-VÄRDEN, MODELLER A TLL F.

Vi börjar med att applicera leave-one-out på den ologaritmerade variabeln och får resultat enligt följande tabell:

Modell	Valda variabler	Förklaringsgrad R^2	PRESS	Typiskt prediktionsfel i kronor
Referens	Alla	0,80	56.4063e+004	167,9
Modell A	x13	0,53	9.1034e+004	67,5
Modell B	x1 x13	0,61	8.6305e+004	65,7
Modell C	x1 x3 x9	0,67	9.5812e+004	69,2
Modell D	x1 x2 x3	0,67	8.0096e+004	63,3
Modell E	x1 x3 x9 x11	0,71	9.2122e+004	67,9
Modell F	x1 x3 x7 x9	0,71	9.3362e+004	68,3

TABELL 6.1.1: PRESS-VÄRDEN, MODELLER A-F

Ovan syns en modell med klar fördel när det gäller att förutsäga värden utanför modellen, nämligen Modell D. Nämnvärt är vidare att denna modell var den med högst statistisk signifikans på de inkluderade parametrarna. Detta är dock inte alltför förvånande då en hög signifikans för en given variabel signalerar en statistiskt säker korrelation, vilket medför att de har en sannolikt högre chans att fortfarande förklara variationen när en bit av informationen försvinner.

Modell B uppvisar den näst bästa modellen i avseende på PRESS-värde bland ovanstående modeller och i kombination med dess värde på förklaringsgraden får vi en modell som också bör anses som relativt tillförlitlig.

Sist bör det nämnas att vi ser att referens-modellen med samtliga variabler inkluderade lider av over-fitting, d.v.s. att antalet förklarande variabler i förhållande till antalet observationen medför att det finns en för stor frihet i anpassningen av beta-värden i förhållande till kravet på låg kvadratsumma vilket leder till en modell som förvisso får en hög förklaringsgrad, men bara så länge den underliggande informationen är exakt den givna.

6.2 PRESS-VÄRDEN, MODELLER G TLL J.

Vi övergår nu till att undersöka hur de modellerna som hade en logaritmerad responsvariabel ter sig under korsvalidering. Nedan visas analogt med föregående kapitel resultatet, först ut i form av en graf.

Modell	Valda variabler	Förklaringsgrad R^2	PRESS	$e^{\sqrt{PRESS/n}}$
Referens	Alla	0,76	3,0774	1,4803
Modell G	x7	0,48	0,5027	1,1718
Modell H	x1 x7	0,55	0,4613	1,1640
Modell I	x1 x3 x9	0,64	0,4953	1,1704
Modell J	x1 x3 x7 x9	0,68	0,4767	1,1669

TABELL 6.2.1: PRESS-VÄRDEN, MODELLER G-J

Kort kan vi nämna att vi ser att referens-modellen även efter denna transformering lider av over-fitting, inte helt oväntat.

I övrigt kan vi lägga märke till att samtliga modeller ovan ger väldigt snarlika typiska prediktionsfelskvoter. Vidare märker vi att modellen med näst bäst PRESS-värde har en betydligt högre förklaringsgrad än modellen med bäst sådant värde varför det är möjligt att den är att föredra.

Samtliga av ovanstående modeller ger ett prediktionsfel på ca 17 %, och i ett underlag där medelvärdet på responsvariabeln är 470,6 kronor svarar det mot i snitt 80.0 kronor, det typiska prediktionsfelet blir dock grövre än så när priset är över medel och mindre när priset är lägre än detsamma.

7. RESULTAT OCH SLUTSATS

Efter att ha undersökt modeller med både logaritmerad samt o-logaritmerad responsvariabel kan vi konstatera att det inte verkar finnas oerhört mycket att vinna på att gå över till logaritmerat slutpris när vi anser att modellernas förmåga att prediktera värden utanför underlaget är av vikt. Med ett typiskt prediktionsfel på ca 17% eller 80,0 kr. förefaller det att de föregående modellerna verkar bättre lämpade då inget av deras typiska prediktionsfel överstiger ens 70 kr.

Bland modellerna A-F är det svårt att utse en klar ”vinnare” av de två modellerna med lägst PRESS (B resp. D), då den modell med färre variabler möjligtvis har det enkom som en följd av att en variabel kan förklaras med hjälp av de två som senare istället inkluderas.

Jag anser att vi kan se båda modellerna som legitima i prediktions syfte och använda den modell man har möjlighet till, har man möjlighet att använda båda kan man göra det och jämföra resultaten.

I ett försök att göra valet lättare utfördes en analys på beta-parametrarnas värde under leave-one-out validering, (som en sorts bootstrap) för att se om någon av de två modellerna hade stabilare parameterskattning än den andra medan underlaget ändrades. Resultatet följer nedan.

Modell	Kod	Koefficient	Medel koefficient, leave-one -out	Standardavvikelse, leave-one-out	$\left\lvert\frac{Standardavvikelse}{medel}\right\rvert$
Modell B	x1	0,14	0.1437	0.0196	0.1366
	x13	18,19	18.0499	1.7772	0.0985
Modell D	x1	0,21	0.2132	0.0154	0.0720
	x2	0,16	0.1581	0.0197	0.1247
	x3	-0,20	-0.1955	0.0172	0.0882

TABELL 7.1: MEDELVÄRDEN OCH STANDARDAVVIKELSER UNDER LEAVE ONE OUT

Ovan kan vi observera att det återigen inte finns någon klar vinnare varpå jag återgår till att hävda att båda modellerna kan och bör anses legitima som prediktionsverktyg.

Om vi ska tolka det inledande tecknet på våra variabler, som varit samma i samtliga modeller A-J, så kan vi börja med att observera med att vi kan utläsa att avstånd närmaste konkurrent (x1) alltid skattats som positivt, dvs ett längre avstånd till konkurrenter korrelerar med ett högre pris. Vidare ser vi något som rent intuitivt inte kändes självklart under analysen, nämligen att avstånd till kollektivtrafiken (x2) korrelerar med ett högre pris. Detta betyder således att modellen skattar (med relativt hög signifikans jämte andra variabler i underlaget) priset hos en frisör nära tunnelbanan lägre. Eventuellt kan det ha något med status att göra; att fastigheter vid tunnelbanan anses mindre attraktiva.

Den tredje förklarande variabeln, x3, skattas under alla modeller som negativ vilket är lätt att förstå intuitivt; ett mindre avstånd till köpcenter och tillika folk som med en högre sannolikhet spenderar pengar en given dag korrelerar med möjliggörandet av högre priser hos närliggande frisörer. Sist nämner vi den trettonde variabeln (x13) som stod för antalet kurser per år som korrelerar med ett högre slutpris.

För samtliga av de ovan nämnda variablerna och deras parametrar är det mycket viktigt att påpeka att korrelation inte förutsätter kausalitet, vilket i vardagsspråk ska tolkas som att det inte finns någon garanti för att ett givet pris skulle höjas då en frisör t.ex. bestämmer sig för att gå på dubbelt så många kurser ett år. Däremot absolut ej heller sagt att det är omöjligt att kausalitet skulle kunna föreligga varför det åtminstone förhoppningsvis vore intressant för frisörer att se över parameterskattningarna.

8. DISKUSSION

Som alltid, när vi behandlar matematisk statistik, finns det ett mått av osäkerhet av den utförda analysen. Kort kan vi nämna att det är ovanligt, till den grad att skribenten blev förvånad, att en logaritmerad responsvariabel inte gav ett bättre resultat än det gjorde, då det ofta är att föredra när det handlar om pengar. Överhuvudtaget föder valet av enbart linjära modeller givetvis en begränsning när det gäller förmågan att fånga alla möjliga verklighetstolkningar. Å andra sidan är poängen med matematisk statistik i sin tillämpning ofta att fånga en *accepterbar* modell av verkligheten, vilket jag anser att uppsatsen lyckas med.

Hade det funnits oändligt med tid samt ett större underlag hade jag oerhört gärna vilja gå till botten med att hitta en förklaring till varför de två observationerna med lägst pris bland de inkluderade observationerna konsekvent överskattas av samtliga modeller. Det är möjligt att det rör sig om en effekt som inte fångats upp av det inledande frågeformuläret till frisörerna, det är lika möjligt att det handlar om en variabeltransformation bland de förklarande variablerna som aldrig realiserades för att verkligheten skulle kunnat återspeglas på bästa möjliga sätt.

I linje med att vi har två observationer alltid överskattas sprids också residualerna med ett skevt mönster på många plottar, vilket återigen föreslår att någon variabeltransformation eller information saknas.

9. REFERNSER

Gut Allan, An Intermidiate Course in probability.

Sundberg Rolf, Kompendium I Tillämpad Matematisk Statistik

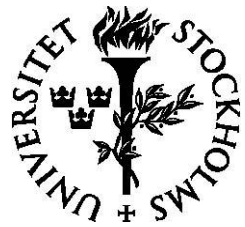
Wikipedia.org (Cross Validation, Leave-one-Out)

SAS Help Section (statistical regression algorithms)

Blom och Holmquist, Statistikteori med tillämpningar

A. APPENDIX

A1: FRÅGEFORMULÄR



Frågeformulär

1. Kontaktinformation

1.1 Företagsnamn

_____

1.2 Adress

_____

_____

_____

_____

1.3 Kontaktperson

Namn: _____

Telefon: _____

Email: _____

2. Frågor för Analys

2.1 Pris på klippning, enkel.

2.2 Närmaste frisör som anses konkurera

2.3 Närmaste kollektivtrafiksanslutning

2.4 Närmaste köpcentrum/handelsgata

2.5 Kontinuerlig eller sekventiell renovering

2.6 Om sekventiell; Datum för senaste renovering

2.7 Antal frisörer i närområdet

2.8 Inriktning [ ] Män [ ]Kvinnor [ ] Båda

2.9 Antal år i samma lokal/område

2.10 Antal år det registrerade företagsnamnet funnits

2.11 Antal Anställda samt Hyrstolar

2.12 Genomsnittlig ålder, anställda

2.13 Genomsnittlig yrkeslivserfarenhet som frisör, anställda

2.14 Antal kurser per år

2.15 Yt-areal för kunder

---

3. Preferenser

3.1 Vill ha den slutliga rapporten skickad per [ ] Post [ ] E-post

3.2 Det går bra att ringa vid eventuella nya frågor [ ] Ja [ ] Helst inte [ ] Nej