NORGES HANDELSHØYSKOLE
Bergen, høsten 2005

Utredning i fordypningsområdet: Finansiering og finansiell økonomi
Veileder: Svein-Arne Persson

PRISING OG ANALYSE AV

BOLIGEIENDOMMER

I BERGEN, ÅSANE OG YTRE ARNA

Av

Hans Leithe

Sammendrag

I denne utredningen har jeg benyttet multippel regresjon og ”hedonic pricing” for å lage en modell for prising av boliger. Målet var å lage en modell for området Arna, Åsane og Bergen sentrum. Ved bruk av dette området forventet jeg å se et mønster i hvordan boligprisen endret seg fra utkant strøk og inn mot sentrum.

I denne sammenhengen har jeg samlet inn data fra Eiendomsverdi sin database på nettet samt studert boligenes ulike egenskaper gjennom salgsprospekter fra www.finn.no. De innsamlede dataene beskriver en rekke karakteristika ved boligene.

Dataene er benyttet til å kjøre en regresjon ved hjelp av ”minste kvadraters metode”. I regresjonen er prisen en endogen variabel, det vil si bestemt innenfor modellen. Prisen fungerer som en funksjon av en rekke eksogene variabler, det vil si bestemt utenfor modellen. For å kunne bruke ”minste kvadraters metode” er det fem antagelser som må være oppfylt. Gjennom analysen har det blitt gjort rede for om modellen har fylt dette kravet. Videre har de ulike karakteristika ved boligen blitt drøftet.

Resultatet av utredningen er en modell som estimerer boligprisen for Arna, Åsane og Bergen gjennom input av de ulike karakteristika ved boligen.

INNHOLDSFORTEGNELSE

Forord

Ved valg av tema for den skriftlige utredningen ønsket jeg å finne et emne som falt innenfor finansiell analyse. I vår semesteret tok jeg blant annet faget ”Metode for finansiell analyse”. Her kom jeg innom temaet ”hedonic pricing” og estimering av blant annet eiendom. Da kom jeg fram til at det å lage en modell for verdiberegning ved regresjonsanalyse ville være en interessant oppgave. Etter noe tankevirksomhet ble verdivurdering av boligeiendommer ved bruk av regresjonsanalyse mitt utgangspunkt for min siviløkonom utredning. Eiendom er et område jeg interesserer meg for og ser for meg at det er et område jeg vil jobbe innenfor etter hvert. Dessuten vil jeg selv snart måtte ut på boligmarkedet etter endt studier og kan ha nytte av utredningen ved å få et bedre innblikk i hvordan boligmarkedet fungere. Modellen er da ment å kunne brukes til å gi et raskt og enkelt estimat på boligprisen både for selger og kjøper.

Siden jeg satser på å komme meg til Oslo gryta etter endt utdanning kunne det vært naturlig å lage en modell for Oslo. Likevel har jeg valg å ta for meg Bergensområdet da jeg ved arbeidet med oppgaven vil befinne meg i Bergen. Dessuten er Bergen og Oslo ikke så ulike med tanke på at det er stort press på boligmarkedet i disse to byene.

Innsamlingen av nødvendig data kunne ha blitt en meget strevsom oppgave hadde det ikke vært for Eiendomsverdi sin database. Vil sende en stor takk til Espen Relbo som ga meg tilgang til basen i forbindelse med utredningen.

Til slutt vil jeg takke min veileder Per-Arne Persson for hjelpen i arbeidet med utredningen.

Bergen, desember 2006

Hans Leithe

1. Problemstilling

Hovedmålet med oppgaven har vært å komme fram til en modell for prising av boliger i området Arna, Åsane og Bergen sentrum. Modellen er ment å kunne gi et raskt estimat på dagens verdi av boligen ved input av ulike karakteristika ved boligen. Estimering av boligpriser langt frem i tid er ikke modellen ment å kunne brukes til. I min oppgave har jeg da ikke sett på eller tatt hensyn til viktige makroøkonomiske faktorer som rentenivå, prisstigning, arbeidsledighet osv. Det er klart at konjunktursvingninger påvirker antall boliger som blir solgt og da prisen gjennom tilbud og etterspørsel. Salgsoppgavene er hovedskapelig fra år 2005 men er supplert med boliger solgt på slutten av år 2004. Det skal da ikke være vesentlige forskjeller i datagrunnlaget pga makroøkonomiske faktorer.
Målet var ikke bare å komme fram til en modell som raskt kunne gi et estimat av boligprisen. Et viktig poeng med oppgaven var å komme fram til en modell som ga best mulig riktig estimat. Gjennom å følge statistisk teori og dens lover skulle jeg komme fram til en slik modell ved bruk av det innsamlede datautvalget.
Som en tilleggsoppgave til slutt brukte jeg datautvalget til å se om det var stor forskjell på meglernes prisantydning og den endelige salgsprisen.

2. Statistisk teori

2.1 Innledning

Før jeg går inn på selve analysen vil jeg kort gå gjennom teorien som ligger bak. Hva slags type regresjon jeg benytter, hvordan man skal tolke regresjonen og de ulike variablene, og hvilke problemer som kan oppstå ved denne type regresjon.
I min oppgave skal jeg forsøke å kunne predikere salgsprisen på en bolig ut fra boligens karakteristika.

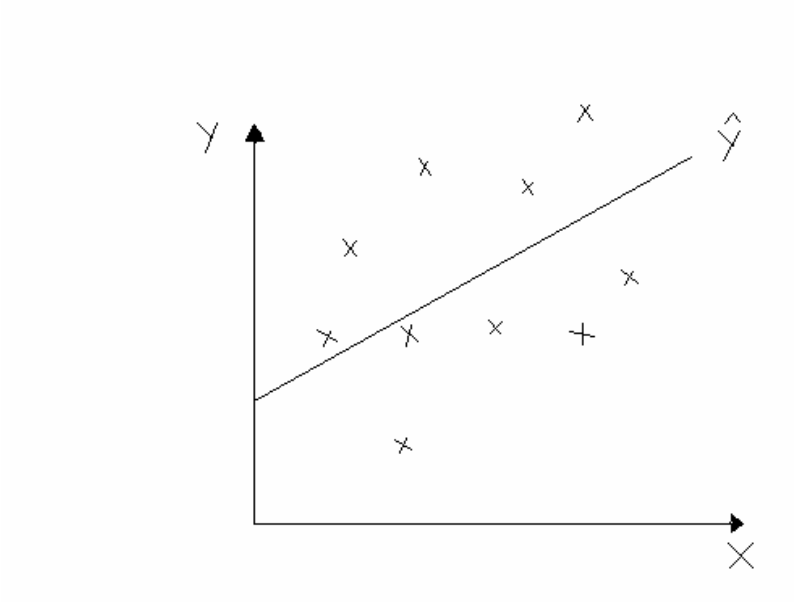
2.2 Regresjonstype – multippel regresjon og "hedonic pricing"

Kort forklart er regresjon en måte å forklare hvordan en avhengig variabel kan forklares ut fra en eller flere uavhengige variabler. Den avhengige variabelen er den vi ønsker å si noe om, mens de uavhengige variablene skal forklare den avhengige variabelen (Brooks 2004).
En enkel regresjon inneholder bare en uavhengig variabel som skal forklare den avhengige variabelen. I min regresjon benytter jeg meg av ”Multippel regresjon” hvor den avhengige variabelen, salgsprisen, forklares ut fra karakteristika ved boligen som er uavhengige variabler. Enkel regresjon kan utføres ved matematiske formler som kan regnes for hånd, mens multippel regresjon er verre å regne for hånd men kan enkelt utføres ved hjelp av dataprogram. Eviews er et slikt dataprogram som jeg har benyttet meg av i analysen.

I den multiple regresjonen har jeg benyttet meg av ”hedonic pricing” som benyttes for å verdsette real aktiva som blant annet eiendom og bolig (Brooks 2004). ”Hedonic pricing” vil si at de ulike karakteristika ved boligen blir gjort om til målbare priser og enheter slik at boligprisen for like eller ulike boliger kan predikeres (Rosen 1974). Typisk for denne metoden er at man må ta standpunkt til hvilke variabler som skal benyttes og hvordan man skal gi dem verdi. Dette er nettopp noe jeg har måttet ta standpunkt til i min modell. Siden boliger varierer mye i hensyn til størrelse, beliggenhet og utforming, bør slike forhold tas eksplisitt med i beregningen av salgsprisen.
Oppgaven med ”hedonic pricing” er altså å avgjøre hvilke variabler som skal være med og hvordan enn kan få med en variabel som en vet har betydning. For eksempel gir god utsikt kontra dårlig utsikt en høyere boligpris. Denne variabelen er vanskelig eller i alle fall tidkrevende å få med da man ville måtte observere hver enkelt bolig og gradere utsikten på en eller annen måte.

2.3 Metode

Metoden som jeg bruker i estimeringen av regresjonen, kalles ”Minste kvadraters metode” (Brooks 2004). Denne metoden benytter seg av tidligere observasjoner for å estimere en verdi i fremtiden. For å forklare metoden videre kan det sees med utgangspunkt i figur 2.3 nedenfor.



Figur 2.3

Kryssene på figuren er de observerte verdiene. Ut fra disse kryssene estimeres linjen ved å minimere summen av de kvadrerte avvikene. Det vil si å minimere summen av avstanden mellom kryssene og den estimerte linja. Hvis dette ikke er tilfelle, vil modellen min inneholde mange feil. Feilleddet μ_t representerer mulige feil i modellen. Som figuren viser, estimeres en lineær regresjon (rett linje). Dermed antar jeg at relasjonen mellom prisen på boligen og de ulike karakteristikkene på huset er på lineær form.

2.4 Forutsettinger for modellen

Ved bruk av Minste kvadraters metode må fem forutsetninger oppfylles som tar utgangspunkt i feilleddet μ_t. Blir ikke disse forutsetningene oppfylt kan det føre til feil resultat ved modellen. De fem forutsetningene er (Brooks 2004):

1. $E(\mu_t)=0$	Feilleddene har forventning lik null. Det vil si at avviket mellom det observerte og det estimerte er lik null.
2. $Var(\mu_t)=\sigma^2 <\infty$	Variansen til feilleddet er konstant og uendelig for alle verdier.
3. $Cov(\mu_i,\mu_j)=0$	Feilleddene er uavhengige av hverandre.

4. $Cov(\mu_t, x_t)=0$ Det er ikke noen sammenheng mellom feilleddene og tilhørende x-verdi

5. $\mu_t \sim N(0, \sigma^2)$ Feilleddet er normalfordelt.

For at modellen skal bli best mulig er det en forutsetning at disse fem punktene blir holdt.

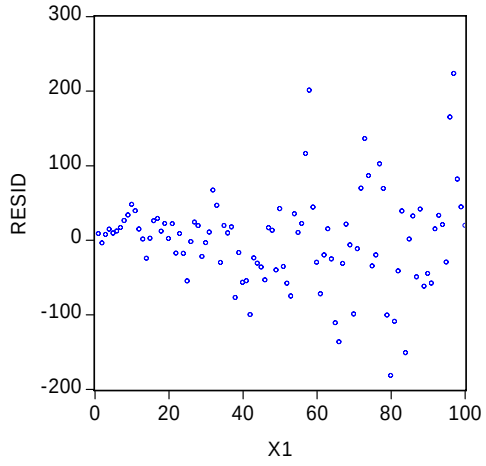
2.5 Problemer som kan oppstå

Ved brud på forutsetningene ovenfor kan dette gi problemer med modellen som må løses for å få en god modell. Dette er problemer som har med heteroskedastisitet (ikke-konstant varians), autokorrelasjon, normalfordelingen og multikolinearitet.

Så lenge regresjonen inneholder et konstantledd vil forventningen til feilleddet bli lik null uansett, så det vil ikke være brudd på 1. antagelse i modellen min (Brooks 2004).

2.5.1 Heteroskedastisitet

En av antagelsene ved minste kvadrats metode er at feilleddene har konstant varians, som kalles homoskedastisitet. Hvis de ikke har konstant varians har vi heteroskedastisitet. En god illustrasjon på heteroskedastisitet eller ikke kan sees ved å plotte feilleddene fra en regresjon mot en av de uavhengige variablene. Figur 2.5.1 nedenfor er et eksempel på heteroskedastisitet hvor vi ser at variansen øker systematisk med variabelen.



Figur 2.5.1

Det som kan skje hvis det er heteroskedastisitet til stede og dette blir oversett er at estimeringen av koeffisientenes standardfeil blir upålitelig og gi feil t-verdi. Grunnen til man kan få brudd på forutsetningen om konstant varians er at det forutsettes at modellen har en lineær sammenheng når det egentlig ikke er det (Brooks 2004).

For å finne ut om det er heteroskedastisitet eller ikke kan man bruke residual plott som ovenfor eller man kan benytte seg av ”White heteroscedasticity test” i Eviews som er en bedre metode for å konstantere heteroskedastisitet eller ikke (Brooks 2004).
For å løse heteroskedastisitet må man forsøke å finne den rette modellen for det datagrunnlaget som skal analyseres. Med det menes at modellen må innholde variabler, både den avhengige og de uavhengige, slik at modellen ikke har heteroskedastisitet. I tillegg kan man se på de enkelte variablene om de inneholder ekstrem verdier som ligger langt fra gjennomsnittet. Hvis så er tilfelle kan det hjelpe og sette variabelen på logaritme form. Ved at den skrives om til den naturlige logaritmen vil ekstreme verdier bli ”dratt” inn slik at avviket ikke blir så stort (Brooks 2004). I min oppgave er vanskelig å løse heteroskedastisitet på noen annen måte.

2.5.2 Autokorrelasjon

En av antagelsene for minste kvadrats metode er at det ikke er korrelasjon mellom feilleddene i modellen. Er det samvariasjon mellom feilleddene har vi det vi kaller autokorrelasjon som er uønsket. Autokorrelasjon vil si at det er en sammenheng mellom feilleddene. Problemene her går ut på det samme som ved heteroskedastisitet. Estimeringen av koeffisientenes standardfeil kan bli gale og man vil få feil t-verdi som kan føre til feil konklusjon av de enkelte variablene (Brooks 2004).
I Eviews kan man teste om man har autokorrelasjon ved bruk av Durbin Watson test eller Breusch-Godfrey test. Durbin Watson test tester om det er sammenheng mellom et feilledd og feilleddets foregående verdi, altså 1.ordens autokorrelasjon. Breusch-Godfrey test kan derimot teste opp for autokorrelasjon opp til n-te orden. Hvis det er autokorrelasjon i modellen må den justeres. Justeringen kan gjøres ved å ta med et nytt ledd, kalt ”hvit støy” som skal justere for autokorrelasjon slik at estimeringen blir mer pålitelig. Det er mest sannsynlig å få problemer med autokorrelasjon når man benytter seg av tidsserie data (Brooks 2004). Derfor forventer jeg ikke å få problemer med det i min oppgave, men må sjekke det likevel.

2.5.3 Normalitet

Ved bruk av minste kvadrats metode er en av antagelsene at feilleddene er normalfordelte. Det er et krav for å kunne utføre hypotese testing av modellens og dens variabler. Tester normalitet ved såkalt ”Bera-Jarque” test i Eviews. For å si noe om hvordan fordelingen til feilleddene er, har man to verdier man ser på, skewness og kurtosis. Skewness sier om fordelingen er symmetrisk i forhold til forventningen mens kurtosis forteller hvor tung halen

til fordelingen er. Det viser seg ofte at det er noen få feilledd som har ekstreme verdier som fører til brudd på normalitet. Løsningen for å oppfylle antagelse om normalfordeling er da å fjerne de boligene som gir ekstreme avvik (Brooks 2004).

2.5.4 Mulitkolinearitet

Ved bruk av minste kvadrats metode er en implisitt antagelse at de uavhengige variablene ikke er korrelerte. Det vil si at koeffisientene til de uavhengige variablene ikke forandres selv om en variabel tilføres eller fjernes fra modellen. Viser det seg at to eller flere uavhengige variabler har sterk samvariasjon, kan et problem kalt multikolinearitet oppstå. Multikolinearitet er når to eller flere variabler er så sterkt korrelert at det er vanskelig å se hvilken variabel som har effekt på den avhengige variabelen. Endrer en variabel seg vil den korrelerte variabelen endre seg tilsvarende. Typisk tegn på multikolinearitet er at t-verdiene ikke er signifikante mens f-testen for regresjonen er signifikant og at resultatgraden er høy. Resultatet er at variablene kan bli feilaktig ikke-signifikante og at de har ulogisk fortegn (Brooks 2004).

Løsning på multikolinearitet:

- Overse problemet er rett og slett en mulighet hvis modellen er tilstrekkelig god nok. Det vil si at koeffisientene har logisk fortegn og er nokså signifikante og regresjonen som helhet er signifikant.
- Fjerne problemet ved å droppe en av de korrelerte variablene. Men hvis de korrelerte variablene begge er av stor teoretisk interesse å ha med kan de vanskelig droppes.
- Øke størrelsen på utvalget kan være en løsning på problemet.
- En mulighet er å lage en rate av de korrelerte variablene og bruke denne i regresjonen. Men igjen kan dette være lite logisk å gjøre da den enkelte variabelen er av stor interesse for regresjonen, å se hvordan den påvirker den avhengige variabelen.

Det er vanskelig å si sikkert om det er problemer med multikolinearitet. En mulighet er å se på en korrelasjonsmatrise mellom variablene. De variablene som har høyest korrelasjon, gjelder enten de er negativ eller positive, er de man kan forvente kan gi problemer med multikolinearitet.

3. Hypotesetesting

Vil her gå gjennom hvilke hypotese tester jeg gjennomfører i oppgaven og hvilke tall jeg studerer for å kunne si noe om testen.

3.1 Testing av regresjonen

Ved testing av selve regresjonen ser jeg på F-verdi, t-verdier og forklaringsgrad.
F-verdi sier om alle variablene sett under et, altså hele modellen, er signifikant forskjellig fra null som er et nødvendig resultat for å si at modellen forklarer noe som helst. Blir verdien ikke-signifikant er modellen ubrukelig (Brooks 2004). Tester på 5 % nivå.
For å se på hvor bra hver enkelt variabel passer inn i modellen ser vi på variablenes t-verdi. Den sier om en variabel er signifikant forskjellig fra null som er ønskelig for å si at variabelen har forklaringskraft i modellen. Er en variabel ikke-signifikant, altså ikke forskjellig fra null, droppes den fra modellen.
Forklaringsgraden ($R^{2)}$ sier noe om hvor mye de uavhengige variablene forklarer den avhengige variabelen. Det er ønskelig å få en høyest mulig forklaringsgrad. I modellen vil det være mest hensiktmessig å se på den justerte forklaringsgraden. Forklaringsgraden R^2 vil alltid øke eller være minst like stor når modellen tilføres en ny variabel. Justert forklaringsgrad justerer derimot for økningen av antall forklarende variabler slik at man kan se hvilke variabler som tilfører modellen forklaringskraft. Øker ikke justert forklaringsgrad ved å tilføre en variabel droppes den fra modellen.

3.2 Testing av heteroskedastisitet

Ved testing av heteroskedastisitet er nullhypotesen homoskedastisitet. Forkastes H_0 har man heteroskedastisitet. Test verdien fåes ved å multiplisere forklaringsgraden R^2 med antall observasjoner i testen og sammenligne den mot kritiske verdi fra tabell. I Eviews får man fram test verdien med tilhørende p-verdi slik at man bare trenger å kunne tyde resultatene. Er p-verdien lavere enn 0,05 har vi heteroskedastisitet på 5 % nivå.

3.3 Testing av autokorrelasjon

Testing kan som sagt foretas med Durbin-Watson test ved et lagg. Det betyr at man tester om det er korrelasjon mellom feilleddet μ_t og feilleddet μ_{t-1} et steg tilbake i tid. Men har i oppgaven brukt Breusch-Godfrey test hvor man kan sjekke for autokorrelasjon for så mange lagg man ønsker. Test verdien fåes også her ved å multiplisere forklaringsgraden med antall

observasjoner som sammenlignes mot kritisk verdi fra tabell. Men disse tallene er bare å lese av i utskrift fra Eviews. Er p-verdien lavere enn 0,05 har vi autokorrelasjon på 5 % nivå.

3.4 Testing av normalfordeling

Testing av normalfordeling foregår ved å se om feilleddene som kommer ut av regresjonen er normalfordelt eller ikke. I Eviews får man ut test-verdien Jarque-Bera med tilhørende p-verdi. Nullhypotesen er at feilleddene er normalfordelt. Er p-verdien høyere enn 0,05 godtas H_0 og feilleddene er normalfordelt.

4. Beskrivelse av datautvalget

Innhenting av data var en tidkrevende oppgave. Jeg var i det hele tatt ikke sikker på om jeg ville få tak i gode nok data til å gjennomføre oppgaven. Hovedproblemet var å få tak i solgte salgsprospekter med salgsprisen. Men takket være tilgang til www.eiendomsverdi.no sin database fikk jeg full tilgang til solgte boliger med de nødvendige opplysningene. Ved hjelp av kartboken i gule sider fant jeg gateadresser som jeg søkte på.

Som sagt valgte jeg å se på boliger i området fra Arna, Åsane til Bergen sentrum. Mitt datautvalg på 200 solgte boliger er tatt fra perioden desember 2004 til september 2005. De fordeler seg med 87 boliger fra Bergensområdet, 63 fra Åsane området og 50 fra Arna området. Jeg antar at dette skal være et godt nok utvalg for å komme fram til en god modell.

Hvilke variabler jeg valgte å ta med i datautvalget kom jeg fram til blant annet ved å ha snakket med eiendomsmeglere og ved å kikke på hvilke opplysninger som gikk igjen i salgsprospektene. Da jeg snakket med "Knut Meeg Torbjørnsen" i Notar om hvilke variabler som hadde størst betydning for boligprisen, svarte han; "Det er tre faktorer som er avgjørende for prisen. Det er beliggenhet, beliggenhet og beliggenhet". Utsagnet understreker viktigheten av boligens beliggenhet. Variabler som boligens avstand til sentrum og nærmeste kjøpesenter var viktige faktorer i følge Torbjørnsen. Med tanke på beliggenhet kunne jeg tatt med flere variabler, men det var viktig å begrense det til de aller viktigste og de som var mulig å ta med. For eksempel en variabel som er vanskelig å få med, men som har mye å si på boligprisen, er utsikt. En bolig med fin vakker utsikt forventes å koste mer enn en bolig bortgjemt bak høye trær. Jeg kunne ha laget en dummyvariabel (forklaring s. 18) her med 1 for god utsikt og 0 for dårlig utsikt. Dette lot seg i midlertidig ikke gjennomføre da jeg fysisk ville ha måttet sjekke hver bolig for så å komme med en subjektiv mening om utsikten. At utsikt ikke kom med i modellen vil jeg påstå er en svakhet ved modellen.

De variablene jeg valgte å ta med framkom alle fra salgsprospektene. Jeg endte opp med hele 19 variabler som jeg valgte å ta med i den første regresjonen. Forklaring på hva variabelen står for kommer senere i oppgaven sammen med hypotesene for variablene.

Ramser opp de 19 variablene nedenfor:

- BOA (boareal)
- Alder
- Soverom
- Balkong
- Bredbånd
- ”BTA” (tilleggsareal)
- Eiendomstype
- Gjeld
- Fliser
- Garasje
- Hage
- Kabel-TV
- Avstand til Bergen sentrum
- Avstand til nærmeste kjøpesenter
- Oppusning
- Parkett
- Tomt
- Varmekabler

Beskrivende data av datautvalget er lagt ved som vedlegg 1.

5. Presentasjon av regresjonsmodellen

P_i = salgspris for boligen

BOA = boareal gitt ved antall kvadratmeter

lnA = den naturlige logaritmen til alderen, (e ≈ 2,718 som grunntall)

Bad = antall bad, (har verdi 1 for de som har flere enn et bad, 0 for et bad)

S = antall soverom, (har verdi for de som har 2 eller flere soverom, 0 for et soverom)

B = balkong, (1 for balkong, 0 for ikke balkong)

BB = bredbånd, (1 for bredbånd, 0 for ikke)

BTA = tilleggsareal gitt ved antall kvadratmeter, (BTA = bruttoareal minus boligareal)

E = eiendomstype, (0 for leilighet, 1 ellers)

Gj = gjeld på boligen

F = fliser på bad

G = garasje, (1 for garasje, 0 ellers)

H = hage, (1 for hage, 0 ellers)

KTV = kabeltv, (1 for kabeltv, 0 ellers)

K_1 = avstand til Bergen sentrum

K_2 = avstand til nærmeste kjøpesenter

O = oppussing, (1 for oppussing, 0 ellers)

Pa = parkett, (1 for parkett, 0 ellers)

T = tomt

V = varmekabler på bad, (1 for varmekabler, 0 ellers)

μ = hvit støy

Dette gir regresjonen:

$$P_i = \beta_0 + \beta_1 BOA + \beta_2 A + \beta_3 Bad + \beta_4 S + \beta_5 B + \beta_6 BB + \beta_7 BTA + \beta_8 E + \beta_9 Gj + \beta_{10} F + \beta_{11} G + \beta_{12} H + \beta_{13} KTV + \beta_{14} K_1 + \beta_{15} K_2 + \beta_{16} O + \beta_{17} Pa + \beta_{18} T + \beta_{19} V + \mu$$

5.1 Beskrivelse av variablene i regresjonen

Nedenfor kommer en beskrivelse av alle variablene som er tatt med i datautvalget og som ble benyttet i den første regresjonen. Variablene er forklart i sammenheng med hypotesene om hvilke fortegn jeg forventer de skal få. Målet her var å se hvilke variabler som ville gi en

signifikant betydning på salgsprisen. Nullhypotesen H_0 er at variabelen ikke har noe å si for modellen. Det vil si at koeffisienten til variabelen er lik null. Alternativhypotesen H_1 er derimot at variabelen har noe å si for modellen, at koeffisienten til variabelen er forskjellig fra null, enten positiv eller negativ betydning for størrelsen på boligprisen. Har bare satt opp alternativhypotesen for variablene da nullhypotesen er null.

En liten forklaring på notasjonene nedenfor. β står for koeffisienten til variabelen, og uttrykket "$\beta>0$" betyr at koeffisienten til variabelen forventes å bli større enn 0**.** Dummy variabler er variabler som enten får verdien eller 0. Et eksempel er balkong som får verdi 1 for balkong og 0 ellers. Hypotesen blir da satt opp på formen; Balkong(B): H_1: $P(B=0)<P(B=1)$. Uttrykket betyr at 1, altså balkong, gir større verdi enn 0, altså ikke balkong.
I forklaringen nedenfor henviser jeg til de beskrivende data i vedlegg 1 når jeg kommenterer hva forventingen til de enkelte variablene ble.

Boarealet (BOA): H_1: $\beta>0$

BOA er i følge Norsk Standard (NS 3940) "Arealet for boligens hoveddel og inneholder entré/hall, oppholdsrom, soverom, kjøkken, bad, wc, vaskerom o.l." (http://www.nef.no/3672.asp).
BOA er den variabelen som har størst betydning for salgsprisen. Desto høyere BOA, desto større salgspris. Prisen pr m^2 BOA er mest sannsynlig ikke like høy i Arna som i Bergen sentrum. Forskjellen i mitt datautvalg ved å dele salgsprisen på BOA er fra 6383 kr pr m^2 (bolig i Arna) til hele 50 000 kr pr m^2 (bolig i Bergen sentrum). Det virker da urimelig å ikke skille mellom for eksempel sentrum og Arna i min modell. Men poenget er at forskjellen skal fanges opp av andre variabler, og spesielt da avstand til sentrum som jeg beskriver nedenfor. Hypotesen er at salgsprisen øker med boarealet (BOA).

lnalder: H_1: $\beta<0$

Alder på en bolig vil jeg tro har stor betydning på boligpris. Det virker logisk å påstå at nye hus koster mer enn gamle hus. Gamle hus vil ofte ha behov for oppussing og større vedlikeholdskostnader. Alderen på boligene fra datautvalget variert fra 0 til 168 år. Dette er et nokså bredt utvalg og vil kunne tenkes å gi store avvik i regresjonen. Derfor har jeg tatt den naturlige logaritmen til alder for å dempe virkningen av de ekstreme verdiene. En av boligene hadde verdien 0, som ga et problem da man ikke kan ta logaritmen til tallet 0. Valgte da å bare erstatte 0 med 1.

Hypotesen er da at salgsprisen synker jo eldre boligen er.

Antall bad: H_1: β>0

Antar at boliger med mer enn et bad vil få høyere salgspris enn de med kun et da bad da bad er et typisk kostbart rom i en bolig. Dessuten kan to bad vise seg å gi mulighet for å ha en utleie enhet i boligen. Har valgt å bruke dummyvariabel her, hvor et bad får verdi 0 og mer enn et bad gir verdi 1. Av datautvalget hadde 10 % mer enn et bad i boligen.
Hypotesen er da at salgsprisen får et tillegg for mer enn et bad.

Soverom: H_1: β>0

Jeg vil forvente at salgsprisen vil øke med antall soverom til en viss grad, men at den er positivt avtagende. Alle boligene har minst et soverom og er naturlig å gi boliger med et soverom verdi 0, altså at boligen kun får merverdi fra og med 2 soverom og mer. Siden jeg forventer at verdiøkningen vil være avtagende velger jeg å sette verdi 1 for alle boliger med 2 soverom eller mer. Spesielt i sentrum er det stor forskjell på 1 og 2 soverom da dette gir mulighet for utleie. I følge Magnus Dehli, megler ved Pareto eiendomsselskap i Oslo, som jeg har snakket med, kan en bolig i Oslo øke fra 1,4 til 1,8 millioner kroner hvor forskjellen er 1 til 2 soverom gitt omtrent samme boareal. Hvis det er mer enn 2 soverom i en bolig vil denne verdiøkningen antas å bli fanget opp av økende boareal. Hele 72,5 % av boligene hadde mer enn et soverom.
Hypotesen er at salgsprisen øker med et tillegg for boliger med mer enn ett soverom.

Balkong(B): H_1: P(B=0)<P(B=1)

I en artikkel i VG (Jensen 2005) ble det av megler Iver Graf antydet at det å bygge balkong til en kostnad av 100 000 kroner ville øke verdien med 600 000 kroner. Dette gjelder for balkong i leiligheter. Antar da sterkt at det å ha balkong vil øke salgsprisen. Da jeg først gikk gjennom dataene ga jeg alle boliger med balkong, terrasse og veranda verdi 1. Det viste seg at hele 70 % fikk verdi 1. Etter en revurdering valgte jeg å gi kun leiligheter med balkong verdi 1 da verdien i det å komme ut i frisk luft uten å forlate boligen er større her enn i en enebolig. Hvis en enebolig ikke skule ha veranda så har den i de fleste tilfeller hage eller lett tilgang på frisk luft. Etter revurderingen fikk 28,5 % verdi 1. Det kunne vært ønskelig å ha hatt med størrelsen på balkongen, men siden det ikke var oppgitt i noen av salgsprospektene måtte jeg droppe dette.
Hypotesen er at salgsprisen øker der hvor en leilighet har balkong.

Bredbånd(BB): H_1: P(BB=0)<P(BB=1)

Jeg har tatt med en variabel for innlagt bredbånd eller ikke da dette er et gode jeg antar folk setter pris på. Det å få lagt inn bredbånd vil medføre en kostnad. Dessuten har boliger i mer avsindige områder ikke mulighet til å få lagt inn bredbånd enda. De som har bredbånd har fått verdi 1 mens de som ikke har fått verdi 0. I datautvalget mitt hadde 30,5 % innlagt bredbånd. Jeg vil legge til at det her kan være en usikkerhet i datautvalget. De salgsprospektene som ikke sier noe om bredbånd har da fått verdi 0 selv om det kan hende at noen av boligene i virkeligheten kan ha bredbånd.

Hypotesen er at salgsprisen øker med et tillegg der det er bredbånd.

Tilleggsarealet (”BTA”): H_1: β>0

BTA står for bruttoarealet til boligen og er oppgitt i alle salgsprospekter. BTA er arealet av hele boligen inklusiv boder og kjellerrom, men eksklusiv terrasser/balkonger og fellesdeler som boder, vaskerom, trapp og lignende i flermannsboliger. (http://www.nef.no/3672.asp.)

Viktig å legge merke til at tilleggsarealet i min modell er BTA minus BOA.

Antar at det er positivt å ha tilleggsareal som kan benyttes til blant annet lager eller eventuelt kan omgjøres til boligareal.

Hypotesen er at salgsprisen øker med tilleggsarealet.

Eiendomstype(E): H_1: P(E=0)<P(E=1)

Til variabelen, eiendomstype, har jeg benyttet en dummy variabel. Enebolig, rekkehus og småhus har jeg valgt sette i en gruppe(enebolig) som får verdien 1 mens leiligheter får verdien 0. Antar her at man oppnår høyere salgspris for en enebolig enn en leilighet. Enn enebolig er typisk større enn en leilighet og har en merverdi i større ”frihet” i tillegg til at det ofte følger med en tomt med hage. 54 % av boligene ble registrert som enebolig.

Hypotesen er at salgsprisen får et tillegg hvis boligen er en enebolig.

Gjeld: H_1: β<0

I prospektene fra eiendomsmeglerne kan man lese av felles gjeld og felles utgifter. Felles utgifter, der de finnes, er små beløp på noen hundre lapper som dekker utgifter som kabel tv, bredbånd og lignende. Har valgt å ikke å ta med denne variabelen da den ikke vil ha mye å si for prisen da dette er utgifter man normalt ville måtte betale. Derimot vil jeg tro at felles gjeld

som borettslag har tatt opp på beboernes vegne vil ha betydning. Denne varierer fra mine data fra 0 til en sum på 900 000 kr.
Hypotesen for denne variabelen blir derfor at felles gjeld vil ha negativ betydning for salgsprisen. Jo større gjeld jo lavere boligpris. Koeffisienten viser da hvor stor andel av gjelden som blir trukket fra før man får boligprisen.

Fliser(F): H_1: P(F=0)<P(F=1)

Som materiale på golv på bad er fliser uten tvil det som foretrekkes. Dessuten er det et tegn på at boligen hvor det er bad med fliser enten er nytt eller oppusset som gir høyere pris. Det kan tenkes at det vil være en viss samvariasjon mellom fliser og oppusning og alder. Av datautvalget hadde hele 81,5 % av boligene fliser på badet.
Hypotesen blir at salgsprisen får et tillegg der det er fliser på badet.

Garasje(G): H_1: P(G=0)<P(G=1)

Jeg antar at garasje gir boligen merverdi. Det å ha muligheten til å ha bilen trygt under tak hvor den for eksempel lettere beskyttes mot innbrudd og hvor den holdes fri for snø og is om vinteren foretrekkes framfor parkering på gata eller gårdsplass. Boliger som har garasje har fått verdien 1 og de som ikke har fått verdien 0. Jeg har valgt å sette verdien 0 til de som har oppstillingsplass. De fleste boligene har en eller annen form for tilgang på parkeringsplass så jeg antar at nettopp parkering i garasje gir en merverdi. 29 % av boligene har garasje.
Hypotesen er at salgsprisen øker met et tillegg der det er garasje.

Hage(H): H_1: P(H=0)<P(H=1)

Jeg antar at det å ha hage er et gode som gir merverdi. Bolig med hage har fått verdi 1 og de uten hage har fått verdi 0. I salgsprospektene blir det ikke oppgitt hvor stor hagen er, bare om det er hage eller ikke. Jeg tar høyde for at størrelsen til en viss grad blir fanget opp av størrelsen på tomten. Hvor stor verdi en hage vil tilføre en bolig kommer an på en rekke faktorer som ikke er målbare i min oppgave, noe som er en svakhet ved variabelen. Faktorer som størrelse, utsikt, beliggenhet og støy ville hatt mye å si her.
53 % av boligene har fått verdi 1, at det er hage. Dette tallet kunne vært mye høyere hvis jeg hadde gitt verdi 1 til leiligheter som har felles hage med de andre i en blokk eller et borettslag. Har valgt å gi leilighet verdi 0 uansett. Dette kan muligens være en feil tolkning. Det kan

hende jeg burde ha vurdert kun tilgangen på hage eller ikke, men jeg velger å tro at det å ha tilgang til egen hage er av størst verdi.
Hypotesen er at salgsprisen øker med et tillegg der det er hage.

Kabel-TV(K): H_1: P(K=0)<P(K=1)
Kabel-TV er i likhet med bredbånd et gode de fleste setter pris. Dette er typisk for byer, tettsteder og borettslag. Innlagt kabel-tv påstår jeg gir boligen en merverdi. Områder som ikke har tilgang til kabel-tv vil da måtte kjøpe seg egen antenne eller parabol for å få inn ønskede kanaler til en dyr penge. Da jeg gikk gjennom datautvalget så jeg en klar sammenheng mellom kabel-tv og bredbånd. Der det var bredbånd var det som regel kabel-tv. Antar da at disse to variablene antakelig vil få stor samvariasjon som kan gi problemer. 52 % av salgsprospektene oppga at boligen hadde kabel-TV.
Hypotesen er at salgsprisen øker med et tillegg der det er bredbånd.

Avstand til sentrum: H_1: β<0
Antar at boligprisen øker jo nærmere sentrum man bor. Sentrum i min modell er da naturlig nok Bergen sentrum.
Årsaker til at det er dyrere å bo nær sentrum er flere. I sentrum har man det beste utvalget av handels-, kultur- og servicetilbud, og det at det er kort vei til disse godene er ettertraktet. Et annet moment er arbeidsplasser. Mange har sine jobber i eller nær Bergen sentrum, slik at det er tid og penger å spare på reise til jobb ved å bo sentralt nær jobben. Bergen er dessuten en studentby med flere tusen studenter. Noen studenter velger å kjøpe bolig når de flytter hit men antar at de fleste kun leier. Likevel er de med på å øke boligpresset slik at prisene presses opp.
For å finne avstanden til sentrum, har jeg benyttet meg av www.kvasir.no sin karttjeneste hvor man legger inn fra og til adresse og får oppgitt nøyaktig avstand og vei beskrivelse. For de boligene som befinner seg i sentrum, har jeg da satt 0 som verdi. Som sentrum for de andre boligene valgte jeg å bruke adressen Torgallmenningen 1 som parameter for å finne avstanden da alle boliger befinner seg nord og nordøst for Bergen.
Hypotesen er at salgsprisen synker med avstand fra sentrum.

Avstand til nærmeste kjøpesenter: H_1: β<0
Som for avstand til sentrum antar jeg at boligprisen øker jo nærmere man bor et kjøpesenter. I et kjøpesenter har man et bredt utvalg i service- og handelstilbud, slik at man kan få gjort unna mange plikter på kort tid. Boliger med beliggenhet i sentrum har alle kort til alle

handelsfasiliteter og har fått verdien 0 km. For boliger utenfor sentrum er Åsane senter parameteren for avstanden. Åsane senter har stort sett de fleste butikker man trenger. I tillegg er blant annet legevakta lokalisert her. Jeg bruker også her kvasir sin karttjeneste for å finne avstand til Åsane senter.
Hypotesen er at boligprisen synker med avstand fra nærmeste kjøpesenter.

Oppusning(O): H_1: P(O=0)<P(O=1)

Nye hus vil som sagt mest sannsynlig koste mer enn gamle hus. Derimot kan dette gi feil resultat i regresjonen hvis man ikke tar med variabelen oppussing. I prospektene fra datautvalget var mange boliger pusset opp i det siste noe som har gitt vesentlig verdiøkning av boligen. Det viser seg at selv boliger over 100 år kan ha like høy pris som en ny bolig etter oppussing. I min modell har jeg valgt å sette verdi 1 for oppussing og verdi 0 for de som ikke er pusset opp. Det har da blitt en skjønnsvurdering av hvilke boliger som har fått verdi 1 eller 0. De dyreste rommene å pusse opp i en bolig er bad og kjøkken. I salgsprospektene hvor det ble nevnt oppusning, så hadde nettopp begge disse to rommene blitt pusset opp samtidig. Dette gjorde det enklere å skille mellom oppussing eller ikke. Ga da verdi 1 kun for de boliger hvor bad og kjøkken var pusset opp. I datautvalget fikk hele 52,5 % verdi 1, altså pusset opp.
Hypotesen er at oppussede boliger får høyere salgspris.

Parkett(P): H_1: P(Pa=0)<P(Pa=1)

I alle salgsprospektene er golvtype i en bolig oppgitt. Jeg antar at parkett er den golvtypen som er av størst verdi siden parkett er oppgitt i 73,5 % av salgsprospektene. Boliger med parkett har fått verdi 1 og 0 ellers. Jeg er derimot ingen ekspert på bedømming av hvilken golvtype som er å foretrekke fremfor noen andre. Det kan hende at en annen golvtype som er oppgitt i salgsprospektene er av større verdi enn parkett. På grunn av dette er jeg derfor usikker på hva resultatet av denne variabelen vil tilføre regresjonen. Dessuten er det ikke nødvendigvis parkett i alle rom i boligen, men har vagt å sette parkett der det er parkett i stue.
Hypotesen er at salgsprisen får et tillegg der det er parkett.

Tomt: H_1: $\beta>0$

Jeg antar at stor tomt foretrekkes framfor liten tomt. Leiligheter oppgis kun som selveiet tomt, og har da fått verdien 0. De resterende eiendomstypene har derimot alle en tomte verdi. Hvor stor verdi en tomt har vil naturligvis avgjøres av hva man kan få ut av tomten. Dette er noe jeg ikke får tatt hensyn til i modellen og er en svakhet. 60,5 % av boligene har verdi høyere enn 0

for tomte areal. Boligene hadde i gjennomsnitt 321,7 m² tomt. Prisen pr m² tomt er som boligpris pr m^2 antakelig høyere i Bergen sentrum enn på Arna. Dette er en svakhet ved modellen da de ikke er lagd noe skille på dette.
Jeg forventer at hage og tomt vil ha høy korrelasjon, slik at en av dem trolig vil droppes i den endelige modellen.
Hypotesen er at salgsprisen øker med antall kvadratmeter tomt.

Varmekabler(V): H_1: P(V=0)<P(V=1)

Som med fliser ønsker man å ha varmekabler på badet. Varmekabler gir gode tørkemuligheter og øker trivsel gjennom et godt og varmt bad. Hele 84 % av datautvalget hadde varmekabler på bad. Jeg vil da ikke være overrasket hvis det vil være høy samvariasjon mellom varmekabler og fliser. Ved gjennomgang av datautvalget så jeg at disse to variablene hang godt sammen.
Det var ikke bare kun på bad at varmekabler kunne være tilstede. Noen av boligene hadde varmekabler i gangen eller på vaskerom. Jeg har derimot ikke valgt å ta med en egen variabel for dette da det gjaldt få av boligene. Dessuten vurderer jeg som det viktigst å ha varmekabler på bad framfor noe annet sted i en bolig.
Hypotesen blir at salgsprisen øker med et tillegg der det er varmekabler på bad.

Variabler som ikke ble med i modellen

Ved gjennomgang av datautvalget tok jeg med en variabel for fyringstype. Jeg satte verdien 1 for de som brukte elektrisk strøm og 0 ellers. Det viste seg at alle boligene hadde elektrisk fyring, mens noen i tillegg hadde vedovn som alternativ. Siden alle hadde verdien 1 valgte jeg å droppe variabelen da den ikke ville kunne skille boligene fra hverandre. Det kunne ha vært aktuelt med en variabel om boligen har peis eller oljefyring for å se om dette kunne ha hatt effekt på boligprisen.
Som jeg har nevnt tidligere hadde det vært ønskelig å få med en variabel for utsikt, men at dette vanskelig lar seg gjennomføre. I tillegg er det en rekke andre variabler med tanke på beliggenhet som kunne vært interessant å se på. Jeg kunne ha sett på avstand til turterreng, jaktområde, barnehage, skole, flyplass, idrettsanlegg, stranda osv. Men jeg valgte i midlertidig å kun ta med beliggenhets variabler for avstand til sentrum og kjøpesenter da modellen skulle bli enklest mulig.

6. Analysen

Etter å ha gått gjennom variablene i modellen og teoriforklaringer er det tid for selve regresjonen og analysen av modellen. Det finnes flere mulig dataprogram man kan benytte for å kjøre regresjoner og testing av regresjonen. Valget mitt falt naturlig nok på statistikk programmet EViews siden jeg har hatt opplæring i programmet gjennom faget "Metode for finansiell analyse".

Målet er å komme fram til en best mulig modell. Jeg vil garantert støte på problemer som har blitt nevnt i teorien ovenfor. Det kan da bli en avveining mellom det som teoretisk er riktig og hva som er logisk riktig å ha med i modellen. Viktige momenter for modellen er å ha høy forklaringsgrad, signifikante t-verdier og logiske fortegn på koeffisient.
Før jeg kjører regresjonen med alle variablene, vil jeg først gi et estimat på prisantydningen som en funksjon av salgsprisen for å kunne sammenligne modellen med prisantydningen fra meglere på boliger lagt ut for salg i dag.

6.1 Prisantydning som funksjon av salgsprisen

I modellen min valgte jeg å ikke ta med prisantydningen i min modell, selv om den ville vært en god pekepinn på hva salgsprisen ville bli. Korrelasjonen mellom prisantydning og salgspris var så høy som 0,959, slik at det var god grunn til å ikke ta med den som forklarende variabel. Dessuten skal jeg lage en modell som nettopp skal komme fram til en prisantydning. Derfor blir det feil å skulle sette inn prisantydning fra en megler da modellens oppgave nettopp er å komme fram til en prisantydning.

Modellen jeg kommer fram til, har salgspris som avhengig variabel. For å kunne sammenligne modellen med prisantydning på nye boliger satt av eiendomsmeglere i dag, må jeg gjøre om salgspris til prisantydning. Ved å kjøre en regresjon med prisantydning som avhengig variabel og med salgsprisen som forklaringsvariabel vil jeg komme fram til en modell for å gjøre om salgsprisen til prisantyding. Regresjonen framkommer av utskriften (tabell 6.1) fra Eviews nedenfor.

Dependent Variable: PRISANTYDNING
Method: Least Squares
Date: 11/28/05 Time: 13:44
Sample: 1 200
Included observations: 200

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-22646.32	36736.47	-0.616453	0.5383
SALGSPRIS	0.920813	0.019373	47.53094	0.0000

R-squared	0.919420	Mean dependent var	1622350.
Adjusted R-squared	0.919013	S.D. dependent var	612246.2
S.E. of regression	174234.2	Akaike info criterion	26.98414
Sum squared resid	6.01E+12	Schwarz criterion	27.01712
Log likelihood	-2696.414	F-statistic	2259.190
Durbin-Watson stat	1.885016	Prob(F-statistic)	0.000000

Tabell 6.1

Regresjonen har høy forklaringsgrad på nesten 92 % og er signifikant. Det blir da å putte inn den estimerte salgsprisen fra modellen min inn i formelen nedenfor. Formelen bruker jeg senere i oppgaven for å sammenligne dagens prisantydning fra meglere på nye boliger lagt ut for salg i dag.

PRISANTYDNING = -22646.3174 + 0.9208134061*SALGSPRIS

6.2 Valg av avhengig variabel

Det ønskelige var å kunne ha salgspris alene som den avhengige variabelen for enklest mulig å kunne komme fram til nettopp estimert salgspris på boligen. Ellers kunne det tenkes at man ville få en bedre modell ved å ha avhengig variabel som; salgspris pr kvadratmeter, logaritmen til salgspris eller logaritmen til salgspris pr kvadratmeter.

Jeg utførte 4 regresjoner med alle variablene hvor jeg da sammenlignet hver av de fire alternativene for avhengig variabel. Logaritmen til salgsprisen hadde ikke problemer med heteroskedastisitet men ga uventet nok problemer med autokorrelasjon. Salgsprisen pr kvadratmeter hadde ikke autokorrelasjon men fikk problemer med heteroskedastisitet. Logaritmen til salgsprisen pr. kvadratmeter ga veldig godt resultat med tanke på

heteroskedastisitet og autokorrelasjon, altså ingen problemer med disse to. Salgsprisen alene ga heller ingen problemer med autokorrelasjon eller heteroskedastisitet, men hadde ikke like overbevisende resultater som Logaritmen til salgsprisen pr kvadratmeter. Jeg valgte likevel å fortsette med salgsprisen som avhengig variabel selv om den var nr 2 i "rangeringen" av beste avhengige variabel, da det var denne jeg ønsket å benytte for enklest mulig kunne komme fram til salgsprisen på en bolig. I tabell 6.2 har jeg satt inn verdiene jeg fant for de fire regresjonene. Tabellen inneholder antall signifikante variabler, antall variabler med riktig forventet fortegn, justert forklaringsgrad, heteroskedastisitet og autokorrelasjon. Jeg minner på at problemer med heteroskedastisitet og autokorrelasjon er til stede når p-verdien er lavere enn 0,05. Har rangert med tanke på heteroskedastisitet og autokorrelasjon.

Avhengig variabel	Rangering	Signifikante variabler	Antall variabler med riktig fortegn	Signifikant modell	Justert forklaringsgrad	Heterosk.	Autokor.
logsalgspris pr m2	1	8	13	JA	0,847	0,313	0,843
salgspris	2	9	14	JA	0,779	0,087	0,216
logsalgspris	3	9	14	JA	0,782	0,581	0,003
salgspris pr m2	4	8	11	JA	0,818	0,049	0,826

Tabell 6.2

6.3 Regresjonen

Jeg kjører først regresjonen med alle variablene slik de er gitt i datautvalget. Det neste vil bli å analysere og utføre nødvendig tiltak for å forbedre modellens troverdighet for så å komme fram til en endelig modell.

Resultatet av regresjonen med salgsprisen som avhengig variabel framkommer av utskriften (tabell 6.3) fra Eviews på neste side.

Dependent Variable: SALGSPRIS
Method: Least Squares
Date: 11/25/05 Time: 13:36
Sample: 1 200
Included observations: 200

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ANT_BAD	-68532.45	88333.93	-0.775834	0.4389
ANT_SOV	393545.4	72977.70	5.392681	0.0000
BALKONG	127416.2	66632.30	1.912229	0.0574
BOA	9945.989	839.2902	11.85048	0.0000
BREDBAND	-44535.79	60082.14	-0.741248	0.4595
BTA	1010.751	998.1042	1.012671	0.3126
EIENDOMSTYPE	496065.2	94486.34	5.250126	0.0000
FE_GJELD	-0.909807	0.185795	-4.896821	0.0000
FLISER	72610.87	100589.1	0.721856	0.4713
GARASJE	75505.03	56697.82	1.331710	0.1846
HAGE	-371880.3	85934.82	-4.327469	0.0000
KABEL_TV	161866.9	67538.25	2.396670	0.0176
KJ_SENTR	58405.16	11103.84	5.259907	0.0000
LNALDER	-80221.28	29109.29	-2.755865	0.0065
OPPPUSING	23666.73	60014.95	0.394347	0.6938
PARKETT	19364.92	62195.44	0.311356	0.7559
SENTRUM	-93050.31	8244.048	-11.28697	0.0000
TOMT	12.91236	84.09135	0.153552	0.8781
VARMKABLER	-33465.93	111365.7	-0.300505	0.7641
C	1161290.	166348.0	6.981088	0.0000

R-squared	0.800438	Mean dependent var	1786460.
Adjusted R-squared	0.779373	S.D. dependent var	637546.0
S.E. of regression	299461.1	Akaike info criterion	28.15200
Sum squared resid	1.61E+13	Schwarz criterion	28.48183
Log likelihood	-2795.200	F-statistic	37.99875
Durbin-Watson stat	1.682415	Prob(F-statistic)	0.000000

Tabell 6.3

6.4 Tolking

Modellen har en høy forklaringsgrad på 80 % som tyder på en god modell. Og enda viktigere, den justert forklaringsgrad er på nesten hele 78 % som er et bedre mål på hvor god modellen er her da den tar høyde for at det er mange variabler med i modellen.

F-verdien og tilhørende p-verdi har signifikante verdier, altså at koeffisientene er forskjellig fra null. Det betyr at modellen er signifikant som helhet. Derimot er mange av variablene ikke

er signifikante på 5 % nivå. 9 av 19 er signifikante, mens to av variablene ligger helt på grensa. 5 av variablene har ulogiske fortegn, eller motsatt av hva jeg forventet de ville bli. Tabell 6.4 nedenfor gir en oversikt over hvilke variabler som ble signifikant og hadde riktig forventet fortegn i forhold til mine hypoteser.

Variabel	Signifikant	Riktig forventet fortegn
BAD	NEI	NEI
SOVEROM	JA	JA
BALKONG	NEI	JA
BOA	JA	JA
BREDBÅND	NEI	NEI
BTA	NEI	JA
EIENDOMSTYPE	JA	JA
GJELD	JA	JA
FLISER	NEI	JA
GARASJE	NEI	NEI
HAGE	JA	JA
KABEL-TV	JA	JA
KJØPESENTER	JA	NEI
LNALDER	JA	JA
OPPUSNING	NEI	JA
PARKETT	NEI	JA
SENTRUM	JA	JA
TOMT	NEI	JA
VARMEKABLER	NEI	NEI

Tabell 6.4

Resultatene tyder på at det må foretas inngrep og justeringer for å få en god modell. Tidligere i oppgaven sjekket jeg at det ikke var problemer med heteroskedastisitet og autokorrelasjon. Men jeg viser likevel hva resultatene av testene ble nedenfor, og kommenterer hvordan de tolkes for å gi et eksempel på hvordan dette gjøres. I tillegg sjekker jeg om feilleddene er normalfordelte som er en av antagelene for at minste kvadrats metode skal holde. Før jeg går videre har jeg en antakelse om at det største problemet her er multikolinearitet. Ved gjennomgang av datautvalget så er det mye som tyder på at flere av variablene vil ha stor korrelasjon. Som forklart i teorien ovenfor kjennetegnes mulitkolinearitet ved høy forklaringsgrad, t-verdier som ikke er signifikante mens f-testen for hele regresjonen er signifikant. Likevel vil jeg påpeke at resultatene er oppløftende til å ha kjørt regresjonen med alle variablene uten å ha gjort tilpasninger.

6.4.1 Heteroskedastisitet

Del av utskrift (tabell 6.4.1) fra Eviews med testing av heteroskedastisitet.

White Heteroskedasticity Test:

F-statistic	1.472076	Probability	0.076244
Obs*R-squared	36.23160	Probability	0.087491

Tabell 6.4.1

Nullhypotesen er homoskedastisitet, det vil si at variansen til feilleddet er konstant. Jeg ser at Obs*R-squared = 36,2316 noe som er lavere enn kritisk verdi på 38,885 (avlest fra chi-kvadrat fordelingen). Nullhypotesen om homoskedastisitet på 5 % nivå kan dermed ikke forkastes. Egentlig trenger jeg ikke å lese av kritisk verdi da Eviews viser hva p-verdien blir. Som utskriften viser ble den 0,087 som tilsier at nullhypotesen ikke kan forkastes på 5 % nivå. Dermed er det ikke problemer med heteroskedastisitet som er en av forutsetningene for minste kvadraters metode. Men jeg er oppmerksom på at den observert verdien ligger veldig nær grensen for å forkaste null hypotesen om homoskedastisitet. Dermed kan små forandringer i modellen kunne føre til heteroskedastisitet.

6.4.2 Autokorrelasjon

Del av utskrift (tabell 6.4.2) fra Eviews med testing av autokorrelasjon.

Breusch-Godfrey Serial Correlation LM Test:

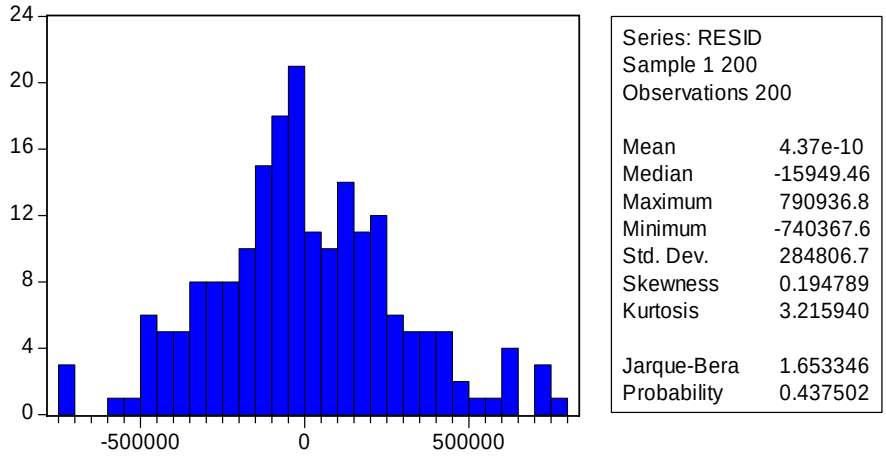
F-statistic	1.281566	Probability	0.273911
Obs*R-squared	7.064558	Probability	0.215883

Tabell 6.4.2

Ser at Obs*R-squared = 7,064 gir p-verdi på 0,216. Nullhypotesen om ikke autokorrelasjon kan dermed ikke forkastes. Altså har vi i følge testen oppfylt nok et krav ved minste kvadrats metode.

6.4.3 Normalitet

Jeg sjekker om feilleddene er normalfordelt ved å kjøre Jarque-Berra test i EViews.



Figur 6.4.3

Testen gir en p-verdi på 0,4375 som er høyere enn forkastningsnivået på 5 %. Dermed kan ikke nullhypotesen om at feilleddene er normalfordelte forkastes. Nok en forutsetning for minste kvadraters metode er da oppfylt. Likevel viser testen at feilleddene har høy verdi for kurtosis, som betyr at fordelingen har en tung hale. Det vil si at det finnes verdier et godt stykke fra normalen, men altså ikke nok til at nullhypotesen om normalfordeling forkastes.
Fra figur 6.4.3 ovenfor ser jeg at det spesielt er tre verdier til venstre og fire verdier til høyre som ligger langt fra normalen. Det kan tenkes at disse observasjonene er årsaken til at noen av variablene ikke fikk signifikante t-verdier og feil fortegn. I Eviews leser jeg av at feilleddene dette gjelder for er for bolig nr 3, 55, 74, 88, 105, 169 og 175 i datautvalget mitt.
Kjørte en ny regresjon uten disse sju observasjonene for å se om modellen ble noe bedre. Resultatet ble ikke helt som håpet. Viste seg at forklaringsgraden økte fra 78 til 81,3 % men at det derimot ble problemer med heteroskedastisitet. Dessuten ble det ingen bedring med tanke på flere signifikante verdier eller forventet fortegn. Konklusjonen er å fortsette med alle observasjonene.

6.4.4 Multikolinearitet

Da gjenstår det å sjekke om vi har problem med multikolinearitet. Tidligere har jeg nevnt at flere av variablene vil kunne ha stor korrelasjon. Jeg sjekker hvordan de enkelte variablene korrelerer med hverandre ved hjelp av EViews.

	Bad	Sov	Balk.	BOA	Bredb.	BTA	Eiendt.	Gjeld	Fliser	Garasje	Hage	K.tv	Kj.sent.	Inalder	Oppus.	Parkett	Sent.	Tomt	Var.k
Bad	1																		
Sov	0,17	1																	
Balk.	-0,14	-0,45	1																
BOA	0,45	0,65	-0,4	1															
Bredb.	-0,15	-0,35	0,26	-0,35	1														
BTA	0,2	0,35	-0,3	0,54	-0,29	1													
Eiendt.	0,17	0,55	-0,68	0,59	-0,33	0,44	1												
Gjeld	-0,02	-0,09	0,08	-0,14	0,16	-0,14	-0,17	1											
Fliser	-0,1	-0,18	0,16	-0,13	0,29	-0,23	-0,21	0,07	1										
Garasje	0,04	0,32	-0,33	0,37	-0,14	0,19	0,48	-0,1	-0,03	1									
Hage	0,15	0,5	-0,54	0,58	-0,38	0,44	0,82	-0,2	-0,32	0,47	1								
K.tv	-0,21	-0,48	0,3	-0,53	0,59	-0,43	-0,44	0,08	0,21	-0,27	-0,5	1							
Kj.sent.	0,08	0,41	-0,31	0,46	-0,32	0,5	0,52	-0,1	-0,2	0,19	0,5	-0,53	1						
Inalder	0,02	0,07	-0,04	-0,05	-0,15	0,16	0,01	-0,3	-0,13	-0,14	0,01	-0,11	-0,03	1					
Oppus.	0,02	0,03	-0,11	-0,07	-0,04	0,05	-0,07	-0,2	0,32	0,14	-0,1	-0,05	-0,13	0,53	1				
Parkett	0,05	-0,22	0,03	-0,08	0,3	-0,23	-0,05	0,06	0,44	-0,02	-0,2	0,31	-0,26	-0,13	0,15	1			
Sent.	0,06	0,51	-0,36	0,54	-0,36	0,47	0,6	-0,1	-0,26	0,3	0,6	-0,6	0,93	-0,1	-0,16	-0,29	1		
Tomt	0,33	0,42	-0,42	0,62	-0,36	0,53	0,61	-0,2	-0,3	0,34	0,67	-0,55	0,53	0,01	-0,16	-0,27	0,59	1	
Varmek.	0,01	-0,09	0,12	-0,06	0,26	-0,25	-0,16	0,09	0,81	-0,02	-0,3	0,18	-0,13	-0,18	0,29	0,51	-0,2	-0,31	1

Tabell 6.4.4 - De største korrelasjonene er merket gult. Negativ korrelasjon er merket med rødt, mens positiv korrelasjon er merket blått.

Ser av tabell 6.4.4 at problemet med multikolinearitet er mindre enn fryktet. Det er vanskelig å sette noen grense for hva som er høy korrelasjon, men høyere enn 0,8 kan i alle fall tyde på problemer med multikolinearitet. Tiltak for å løse mulitkolinearitet er som nevnt i teorien ovenfor å øke datautvalget, droppe en variabel eller slå sammen variabler i en rate. Økning av utvalget er ikke aktuelt da dette er veldig tidkrevende. Dessuten er det vanskelig å finne salgsprospekter på flere boliger uten å måtte gå lengre bak i tid enn jeg allerede har gjort. Da kan det bli nødvendig å vurdere prisstigningen. Jeg vil da i stedet se om noen av variablene bør utelates eller om to variabler eventuelt kan slås sammen i en rate. Før jeg gjør dette vil jeg gi en kort kommentar til de høyest korrelerte variablene.

Spesielt ser jeg at variablene ”Avstand til sentrum” og ”Avstand til nærmeste kjøpesenter” har korrelasjonskoeffisient på hele 0,93. Fra regresjonen ser jeg at variabelen ”Avstand til nærmeste kjøpesenter” har fått ulogisk feil fortegn selv om den var signifikant. Vil se om modellen blir bedre ved å droppe en av dem.
Varmekabler og fliser antok jeg ville være korrelerte noe resultatet på 0,81 viser. Fra regresjonen ser jeg at nettopp begge disse to har fått ikke signifikante verdier samt at fliser har fått feil forventet fortegn. Siden det er stor sannsynlighet for at der den ene er til stede, vil den andre også være til stede. Derfor vil jeg se om modellen blir bedre ved å droppe en av variablene.
Hage og eiendomstype har også som forventet høy korrelasjon (0,82). Det er naturlig å forvente at de fleste eneboliger har hage. Samtidig nevner jeg at det er en viss korrelasjon mellom tomt og hage (0,67), og mellom tomt og eiendomstype (0,61). Det er altså en viss samvariasjon mellom disse tre som ikke er veldig overraskende. Eiendom har både signifikant verdi og forventet fortegn. Hage har signifikant verdi men ikke forventet fortegn. Tomt har ikke signifikant verdi men forventet fortegn.
Kabel-tv og bredbånd har også en viss korrelasjon på 0,59. Det kan her være en usikkerhet i datautvalget. Jeg hadde inntrykk av at det var litt manglede info om tilstedeværelsen til disse to variablene, og da spesielt bredbånd. Ikke alle salgsprospektene var like utfyllende. Noen inneholdt kun det aller viktigste, og således kan disse to variablene ha blitt utelatt i noen tilfeller. Begge variablene er et gode kjøper antas å sette pris på. Så langt i regresjonen er det derimot bare kabel-tv som har signifikant verdi.

6.5 Forbedre modellen

For å lage en bedre modell med flere signifikant verdier, testet jeg om det var mulig å gjøre noe med avstandsvariablene som hadde høyest korrelasjon. Regresjonen ble kjørt på nytt uten variabelen ”Avstand til nærmeste kjøpesenter” siden denne hadde feil forventet fortegn og var høyt korrelert med ”Avstand til sentrum”. Dette ga ikke noe forbedring av modellen. Den justerte forklaringsgraden falt fra 78 til 74,7 % og det ga ikke flere signifikante variabler. Velger da bare å la variabelen stå som den er. En forklaring på at boligprisen vil øke med avstand fra nærmeste kjøpesenter kan være at man ikke vil bo rett ved siden av et kjøpesenter men rett i nærheten. De fleste boligene har ikke lang avstand til et kjøpesenter slik at boliger litt unna et kjøpesenter, men ikke langt unna, kan være mer ettertraktet. Det kan tenkes at det forbindes med mye bråk ved å bo rett ved kjøpesenteret. Boliger som ligger lengst unna et kjøpesenter vil da unaturlig få høy tillegg i prisen som følge av dette, men det justeres mot at

lang avstand fra sentrum trekker prisen ned. Jeg ser at avstand fra sentrum trekker mer ned enn avstand til kjøpesenter trekker opp. Jeg kjørte også en regresjon hvor variabelen "Avstand til sentrum" ble droppet. Det ga en klar forverring av modellen så jeg beholder begge avstandsvariablene.

For å gå videre tok jeg utgangspunkt i kun de variablene som hadde signifikante verdier og kjørte en regresjon med disse. Resultatet ble som utskriften (tabell 6.5.1) nedenfor viser.

Dependent Variable: SALGSPRIS
Method: Least Squares
Date: 11/25/05 Time: 15:14
Sample: 1 200
Included observations: 200

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1306294.	131746.2	9.915226	0.0000
ANT_SOV	367328.6	68448.68	5.366482	0.0000
BOA	10135.13	668.7764	15.15474	0.0000
EIENDOMSTYPE	445679.5	79383.80	5.614237	0.0000
FE_GJELD	-0.973649	0.181470	-5.365328	0.0000
HAGE	-356797.1	78637.37	-4.537247	0.0000
KABEL_TV	135327.5	57850.29	2.339271	0.0204
KJ_SENTR	58027.37	10314.23	5.625950	0.0000
LNALDER	-77949.75	23450.47	-3.324017	0.0011
SENTRUM	-92089.99	7675.028	-11.99865	0.0000

R-squared	0.789998	Mean dependent var	1786460.
Adjusted R-squared	0.780051	S.D. dependent var	637546.0
S.E. of regression	299001.0	Akaike info criterion	28.10299
Sum squared resid	1.70E+13	Schwarz criterion	28.26790
Log likelihood	-2800.299	F-statistic	79.41719
Durbin-Watson stat	1.696521	Prob(F-statistic)	0.000000

Tabell 6.5.1

Variablene og modellen som helhet beholdt sine signifikante verdier samtidig som justert forklaringsgrad holdt seg på 78 %. Jeg testet om denne modellen hadde brudd på forutsetningene ved minste kvadraters metode, og det viste seg dessverre at modellen hadde problemer med heteroskedastisitet.

Oppgaven videre var da å se om noen av de andre variablene ville bli signifikante sammen med variablene i tabell 6.5.1, og om de løse problemet med heteroskedastisitet. Jeg valgte å kjøre regresjonen med variablene fra tabell 6.5.1 mot de resterende variablene hver for seg for å se om noen av variablene ville bli signifikante. Jeg begynte med de med høyest t-verdi først.

Etter hver regresjon sjekket hvordan det gikk med heteroskedastisitet og autokorrelasjon. Resultatene er ført inn i tabell 6.5.2 nedenfor.

Variabel	Justert R2	T-verdi	Prob	Prob heteros.	Prob auto.
Kun de signifikante	0,78	>2	<0,05	0,010485	0,3044
Balkong	0,7833	1,974	0,0498	0,010332	0,1353
Garasje	0,7812	1,401	0,1629	0,0406	0,3064
BTA	0,7798	0,91	0,3637	0,0237	0,284
Bad	0,7804	-1,15	0,2495	0,0382	0,4479
Bredbånd	0,779	-0,31	0,7517	0,01259	0,308
Fliser	0,7805	1,019	0,2355	0,00508	0,3677
Oppussing	0,7797	0,84	0,3967	0,00478	0,2816
Parkett	0,779	0,4	0,688	0,0051	0,299
Varmekabler	0,7794	0,67	0,5019	0,005228	0,3142
Tomt	0,7789	-0,1869	0,8519	0,02356	0,3124

Tabell 6.5.2

Av tabellen er det kun balkong som fikk signifikant verdi, mens alle de andre forble ikke signifikant. Balkong ga også en liten høyere justert forklaringsgrad. Ingen av variablene løste problemet med heteroskedastisitet. Autokorrelasjon ble ikke i noe tilfelle et problem. Ved å følge statistisk teori, skal modellen kun inneholde de variablene som ble signifikante. Dette gir regresjonen:

SALGSPRIS = 1224499.653 + 389491.8348*SOV + 128734.224*BALKONG + 10081.46262*BOA + 529506.6057*EIENDOMSTYPE - 0.9575572699*GJELD - 359455.1151*HAGE + 133587.3998*KABEL_TV + 59653.07845*KJ_SENTR - 77960.8866*LNALDER - 93895.81527*SENTRUM

Modellen står altså igjen med 10 variabler fra utgangspunktet på 19. Da er alle de 10 variablene signifikante og gir en høy justert forklaringsgrad på 78 %. Modellen har en høy F-verdi som tilsier at modellen er god som helhet, at den er signifikant forskjellig fra null. Dessverre viser det seg at modellen har problemer med heteroskedastisitet. Den opprinnelige modellen med alle variablene hadde ikke heteroskedastisitet, men p-verdien på 8,7 % var veldig nær forkastnings nivå på 5 %. Dessuten skulle jeg helst sett at variabler som antall bad, BTA, garasje, oppusning og tomt hadde blitt med i modellen da jeg forventer at disse vil ha en del å si på salgsprisen.

7. Simulering

Jeg testet modellen med de 10 gjenværende variablene ved å bruke salgsprospekter lagt ut på www.finn.no den siste tiden. Data fra 20 salgsprospekter jeg fant der matet jeg inn i modellen min. Resultatene har jeg ført inn i tabell 7 nedenfor. Nr 1-4 er fra Arna, 5-12 er fra Åsane mens 13-20 er fra Bergen sentrum. Det var for øyeblikket ikke lagt ut flere boliger enn fire for salg i Arna området, men mener det skal holde for å få et bra overblikk på hvordan modellen fungerer i praksis. Simuleringen har jeg foretatt ved hjelp av Excel. Salgsprisen og prisantydningen fra simuleringen er ført inn i tabellen. I tillegg er prisantydningen fra salgsprospektene satt av meglerne ført inn i tabellen.

Nr.	Salgspris	Prisantydning	Prisantydning fra megler	% forskjell
1	2147510	1954809	1450000	34,81 %
2	1327274	1199525	990000	21,16 %
3	827643	739458	990000	-25,31 %
4	2375208	2164476	2290000	-5,48 %
5	1917101	1742645	2000000	-12,87 %
6	3132838	2862111	2990000	-4,28 %
7	1104714	994588	890000	11,75 %
8	600053	529890	490000	8,14 %
9	2096193	1907556	1650000	15,61 %
10	1271257	1147944	980000	17,14 %
11	998775	897038	1190000	-24,62 %
12	2414960	2201080	1800000	22,28 %
13	1907862	1734138	1390000	24,76 %
14	1773682	1610583	1590000	1,29 %
15	2077844	1890659	1690000	11,87 %
16	3019220	2757491	2200000	25,34 %
17	1721019	1562091	1350000	15,71 %
18	1484651	1344439	1490000	-9,77 %
19	1409671	1275397	1390000	-8,24 %
20	2054707	1869355	1890000	-1,09 %
	Gjennomsnitt	1619264	1535000	5,91 %
	Standardavvik	620033	571089	16,97 %

Tabell 7

Tabellen tyder på at modellen ikke fungerer så bra som ønsket. Gjennomsnittelig priser modellen 5,91 % høyere enn prisantydningen til eiendomsmeglerne som jeg ikke kan si meg helt fornøyd med. Dette understrekes av de 20 observasjonenes standardavvik på hele 16,97 %. Ser at modellen treffer veldig dårlig som for eksempel bolig nr 1 hvor avviket er på hele

34,81 % og bolig nr 3 hvor avviket er -25,31 %. Jeg må gjøre oppmerksom på at gjennomsnittet og standardavviket på 5,91 % og 16,97 % er for den % vise forskjellen.

Det ønskelige hadde vært at modellens prisantyding hadde gitt det samme som meglernes prisantydning. For å si noe konkret om modellen foretar jeg en test ut fra statistisk teori av utvalget ovenfor for å se om modellen gir samme resultat som meglerne. Benytter meg av en "Two-sample T-test" (Keller og Warrack 2003) hvor nullhypotesen er at det ikke er forskjell mellom prisantydingen til modellen og meglerne.

H_0: $(\mu_1-\mu_2) = 0$

H_1: $(\mu_1-\mu_2) \neq 0$

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad , \quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} \quad , \mathbf{v = n_1 + n_2 - 2}$$

.

$x_1 = 1619264$, $s_1 = 620033$, $n_1 = 20$

$x_2 = 1535000$, $s_2 = 571089$, $n_2 = 20$

Innsetting av tallene i formlene ovenfor gir; $t_{obs} = 0,447$

Leser av kritisk verdi lik ±2,021 i tabellen for t-fordeling med frihetsgrad (v) 38 (Keller og Warrack 2003). Testobservatoren er mindre enn kritisk verdi slik at nullhypotesen ikke kan forkastes. Ut fra hypotese testing kan man ikke forkaste modellen. Med det menes at man ved testing på 5 % nivå ikke kan si at modellen gir forskjellig prisantydning i forhold til prisantydning gitt av meglerne. Derimot må det huskes på at det er brudd på forutsetningene ved bruk av minste kvadrats metode da det er bevist at det eksisterer heteroskedastisitet ved modellen. Problemer med heteroskedastisitet og at avvikene i flere av de 20 observasjonene var høye, gjør at jeg ikke er fornøyd med modellen slik den er.

8. Ny modell

Salgspris som avhengig variabel ga en modell som ikke fylte antagelsene ved minste kvadraters metode. Med tanke på at regresjon med logaritmen til salgsprisen pr kvadratmeter ga gode resultater med hensyn på heteroskedastisitet og autokorrelasjon, vil jeg undersøke om det kan gi en bedre modell. Den første modellen hadde BOA (boarealet) som en forklarende variabel hvor boligprisen økte med 10 081 for hver kvadratmeter. Det vil si at BOA skulle ha en lineær sammenheng på boligprisen. Nettopp det å ha BOA som en forklarende variabel i modellen tror jeg er hovedgrunnen til at problemer med heteroskedastisitet oppstod. I datautvalget varierte prisen pr kvadratmeter fra 6 383 til 50 000 kroner pr kvadratmeter. Ved å trekke BOA inn i den avhengige variabelen tror jeg problemet med heteroskedastisitet blir løst. I tillegg vil ”ekstrem” verdier bli dempet ved å ta logaritmen til den avhengige variabelen.

I tillegg til å forandre den avhengige variabelen innfører jeg de tre nye variablene; Arna, Åsane og Bergen sentrum. Disse blir dummy variabler hvor boligen for eksempel for Arna får verdi 1 hvis den er lokalisert der og 0 ellers.
Jeg setter opp nye hypoteser for denne modellen da den avhengige variabelen blir logaritmen til salgsprisen pr kvadratmeter.

8.1 Presentasjon av den nye regresjonsmodellen

$\log P_i m^2$ = logaritmen til salgsprisen pr kvadratmeter, (log = 10 som grunntall)

Ar = Arna, (1 for Arna og 0 ellers)

Å = Åsane, (1 for Åsane og 0 ellers)

Be = Bergen, (1 for Bergen og 0 ellers)

A = den naturlige logaritmen til alderen, (e ≈ 2,718 som grunntall)

Bad = antall bad, (har verdi 1 for de som har flere enn et bad, 0 for et bad)

S = antall soverom, (har verdi for de som har 2 eller flere soverom, 0 for et soverom)

B = balkong, (1 for balkong, 0 for ikke balkong)

BB = bredbånd, (1 for bredbånd, 0 for ikke)

BTA = tilleggsareal, bruttoareal minus boligareal

E = eiendomstype, (0 for leilighet, 1 ellers)

Gj = gjeld på boligen

F = fliser på bad

G = garasje, (1 for garasje, 0 ellers)

H = hage, (1 for hage, 0 ellers)

KTV = kabeltv, (1 for kabeltv, 0 ellers)

K_1 = avstand til nærmeste kjøpesenter

K_2 = avstand til Bergen sentrum

O = oppussing, (1 for oppussing, 0 ellers)

Pa = parkett, (1 for parkett, 0 ellers)

T = tomt

V = varmekabler på bad, (1 for varmekabler, 0 ellers)

μ = hvit støy

Dette gir regresjonen:

$$\log P_i m^2 = \beta_0 + \beta_1 Ar + \beta_2 Å + \beta_3 Be + \beta_4 A + \beta_5 Bad + \beta_6 S + \beta_7 B + \beta_8 BB + \beta_9 BTA + \beta_{10} E + \beta_{11} Gj + \beta_{12} F + \beta_{13} G + \beta_{14} H + \beta_{15} KTV + \beta_{16} K_1 + \beta_{17} K_2 + \beta_{18} O + \beta_{19} Pa + \beta_{20} T + \beta_{21} V + \mu$$

8.2 Hypoteser for den nye modellen

Hypotesene blir forandret i forhold til den første modellen da den avhengige variablene er forandret til logaritmen til salgsprisen pr kvadratmeter. Jeg setter opp bare alternativhypotesen for modellen her også.

Arna(AR): H_1: P(Ar=0)<P(Ar=1)

Hypotesen er at logaritmen til salgsprisen pr kvadratmeter øker med et tillegg for beliggenhet i Arna området og 0 ellers. Problemet med stedsvariablene blir hvor grensene mellom de tre stedene skal settes. Det må brukes skjønn og sunn fornuft.

Åsane(Å): H_1: P(Å=0)<P(Å=1)

Hypotesen er at logaritmen til salgsprisen pr kvadratmeter øker med et tillegg for beliggenhet i Åsane og 0 ellers.

Bergen sentrum(Be): H_1: P(Be=0)<P(Be=1)

Hypotesen er at logaritmen til salgsprisen pr kvadratmeter øker med tillegg for beliggenhet i Bergen sentrum og 0 ellers.

lnalder: H_1: $\beta<0$

Hypotesen blir som i første modell at logaritmen til salgsprisen pr kvadratmeter synker med alderen på boligen.

Antall bad: H_1: $\beta<0$

Jeg er litt usikker på fortegnet her men forventer motsatt fortegn her i forhold til første modell. Bolig med mer enn et bad er typisk for bolig med stort BOA, og jeg har tatt høyde for at prisen pr kvadratmeter synker med økt BOA. Logaritmen til salgsprisen pr kvadratmeter vil da synke med et tillegg. Jeg bruker samme dummyvariabel hvor ett bad får verdi 0 og mer enn ett bad gir verdi 1.

Soverom: H_1: $\beta<0$

Samme hypotese som for bad. Logaritmen til salgsprisen pr kvadratmeter synker med et tillegg. Soverom får verdi 1 for antall soverom 2 eller mer, og 0 ellers.

Balkong(B): H_1: P(B=0)<P(B=1)

Balkong får verdi 1 der det er tilstede og 0 ellers. Forventer et positivt tillegg der balkong er tilstede. Variabelen gjelder også her kun for eiendomstype leilighet.

Bredbånd(BB): H_1: P(BB=0)<P(BB=1)

Hypotesen er at bredbånd gir et tillegg.

Tilleggsarealet (”BTA”): H_1: β>0

Logaritmen til salgsprisen pr kvadratmeter øker med antall kvadratmeter BTA.

Eiendomstype(E): H_1: P(E=0)>P(E=1)

Hypotesen er at logaritmen til salgsprisen pr kvadratmeter er lavere for enebolig enn for leilighet. Jeg antar at leilighet er typisk mindre boligenheter enn enebolig slik at antagelsen om at prisen pr kvadratmeter synker med økende boareal. Variabelen får verdi 0 for leilighet og 1 ellers.

Gjeld: H_1: β<0

Hypotesen er at utestående gjeld på boligen drar ned prisen med en andel av gjelden.

Fliser(F): H_1: P(F=0)<P(F=1)

Hypotesen er fortsatt at fliser er et positivt gode som gir et tillegg i prisen. Variabelen får verdi 1 for fliser og 0 ellers.

Garasje(G): H_1: P(G=0)<P(G=1)

Hypotesen er at det å ha garasje gir et tillegg i prisen. Variabelen får verdi 1 for garasje og 0 ellers.

Hage(H): H_1: P(H=0)<P(H=1)

Hypotesen er at hage gir et tillegg i prisen. Hage gir verdi 1 og 0 for ikke hage.

Kabel-TV(K): H_1: P(K=0)<P(K=1)

Hypotesen er at kabel-TV gir et tillegg i prisen. Kabel-TV får verdi 1 og 0 ellers.

Avstand til sentrum: H_1: β<0

Jeg antar at boligprisen synker med antall kilometer fra sentrum. Spesielt har jeg sett at datautvalget for prisen pr kvadratmeter er betydelig høyere jo nærmere sentrum boligen er lokalisert. Dette henger sammen med at boliger i sentrum er typisk mindre leiligheter.

Avstand til nærmeste kjøpesenter: H_1: $\beta<0$

Jeg antar det samme som for sentrum at logaritmen til salgsprisen pr kvadratmeter synker med antall kilometer fra nærmeste sentrum. Likevel er jeg oppmerksom på at det ble motsatt i første modell, og at det er stor sjanse for at det samme skjer her.

Oppusning(O): H_1: P(O=0)<P(O=1)

Ny oppuset hus forventes å koste mer enn gamle slitte hus. Derfor er hypotesen at logaritmen til salgsprisen pr kvadratmeter øker med et tillegg for oppusning. Variabelen har verdi 1 for oppusning og 0 ellers.

Parkett(P): H_1: P(=0)<P(=1)

Hypotesen er at parkett gir et tillegg i prisen. Verdi 1 for parkett og 0 ellers.

Tomt: H_1: $\beta>0$

Hypotesen er at prisen øker med antall kvadratmeter tomt.

Varmekabler(V): H_1: P(V=0)<P(V=1)

Hypotesen er et tillegg i prisen der det er varmekabler på bad. Verdi 1 for varmekabler og 0 ellers.

8.3 Analysen av den nye modellen

Resultatet fra regresjonen med logaritmen til salgsprisen pr kvadratmeter med alle variablene er vist i tabell 8.3.a nedenfor. Jeg bemerker at regresjonen ikke inneholder et konstantledd på samme måte som i modell 1. Konstant leddet blir erstattet med variablene Arna, Åsane og Bergen. Det går ikke å kjøre denne regresjonen med et annet konstant ledd i tillegg i Eviews. Får problemer med ”near singular matrix”. Det vil si at de tre variablene er ”mutually exclusive” dummy variabler, slik at de er fullstendig korrelert med konstant leddet. Da må konstant leddet droppes får at disse tre variablene kan bli med i modellen.

Dependent Variable: LOGSALGSPRIS_PRM2
Method: Least Squares
Date: 11/27/05 Time: 21:00
Sample: 1 200
Included observations: 200

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ANT_BAD	-0.062213	0.022056	-2.820702	0.0053
ANT_SOV	-0.071703	0.016953	-4.229484	0.0000
ARNA	4.317322	0.072606	59.46219	0.0000
ASANE	4.394814	0.058760	74.79281	0.0000
BALKONG	0.022889	0.017344	1.319728	0.1886
BERGEN	4.532991	0.042723	106.1027	0.0000
BREDBAND	0.031282	0.015569	2.009304	0.0460
BTA	-0.000527	0.000250	-2.111556	0.0361
EIENDOMSTYPE	0.078760	0.024528	3.211031	0.0016
FE_GJELD	-2.61E-07	4.82E-08	-5.420849	0.0000
FLISER	-0.023772	0.026498	-0.897101	0.3709
GARASJE	-0.003753	0.015136	-0.247972	0.8044
HAGE	-0.065284	0.022386	-2.916295	0.0040
KABEL_TV	0.035824	0.018003	1.989965	0.0481
KJ_SENTR	0.010617	0.004670	2.273350	0.0242
LNALDER	-0.013554	0.007556	-1.793778	0.0745
PARKETT	-0.005821	0.016140	-0.360629	0.7188
OPPUSSING	0.018681	0.015668	1.192315	0.2347
SENTRUM	-0.012743	0.004943	-2.578109	0.0107
TOMT	3.79E-05	2.19E-05	1.726464	0.0860
VARMKABLER	0.010255	0.029244	0.350688	0.7262

R-squared	0.868799	Mean dependent var	4.284549
Adjusted R-squared	0.854139	S.D. dependent var	0.203362
S.E. of regression	0.077667	Akaike info criterion	-2.173692
Sum squared resid	1.079770	Schwarz criterion	-1.827369
Log likelihood	238.3692	Durbin-Watson stat	1.947399

8.3.a

Resultatene fra denne regresjonen har lovende verdier. 13 av variablene er signifikante og ytterlige to ligger helt på grensa. Forklaringsgraden (86,88 %) og justert forklaringsgrad (85,41 %) er begge høyere enn i den forrige modellen. 7 av variablene har feil forventet fortegn som her kan skyldes multikolinearitet eller at modellen ”naturlig” tilpasses for å finne den mest riktige verdien. Oversikt over hvilke variabler som ble signifikant og fikk riktig forventet fortegn i forhold til mine hypoteser sees i tabell 8.3.b nedenfor.

Variabel	Signifikant	Riktig forventet fortegn
BAD	JA	JA
SOVEROM	JA	JA
ARNA	JA	JA
ÅSANE	JA	JA
BALKONG	NEI	JA
BERGEN	JA	JA
BREDBÅND	JA	JA
BTA	JA	NEI
EIENDOMSTYPE	JA	NEI
GJELD	JA	JA
FLISER	NEI	NEI
GARASJE	NEI	NEI
HAGE	JA	NEI
KABEL-TV	JA	JA
KJØPESENTER	JA	NEI
LNALDER	NEI	JA
PARKETT	NEI	NEI
OPPUSNING	NEI	JA
SENTRUM	JA	JA
TOMT	NEI	JA

Tabell 8.3.b

8.3.1 Heteroskedastisitet

Del av utskrift (tabell 8.3.1) fra Eviews med testing av heteroskedastisitet.

White Heteroskedasticity Test:

F-statistic	0.923174	Probability	0.575991
Obs*R-squared	24.36773	Probability	0.554947

Tabell 8.3.1

Nullhypotesen er homoskedastisitet, det vil si konstant varians. Jeg ser at Obs*R-squared = 24,36 noe som er lavere enn kritisk verdi på 38,885 (avlest fra chi-kvadrat fordelingen).

Nullhypotesen om homoskedastisitet på 5 % nivå kan derfor ikke forkastes. Dermed er det ikke problemer med heteroskedastisitet her som er en av forutsetningene for minste kvadraters metode. Jeg kan konstantere at det er mye bedre klarering med tanke på heteroskedastisitet her med p-verdi på 0,555 enn i forhold til den første modellen. Heteroskedastisitet er som sagt tilstede hvis p-verdien er lavere enn 0,05.

8.3.2 Autokorrelasjon

Del av utskrift (tabell 8.3.2) fra Eviews med testing av autokorrelasjon

Breusch-Godfrey Serial Correlation LM Test:

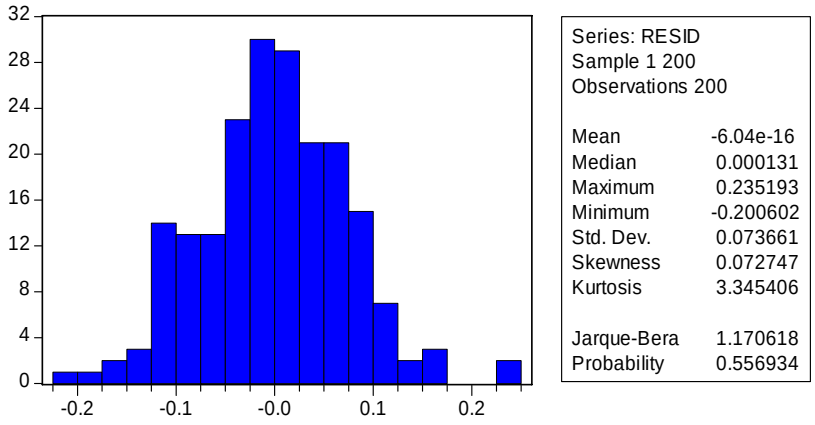
F-statistic	0.021126	Probability	0.999810
Obs*R-squared	0.121338	Probability	0.999739

Tabell 8.3.2

Jeg ser at Obs*R-squared = 0,121 gir p-verdi på hele 0,999. Nullhypotesen om ikke autokorrelasjon kan dermed trygt godtas. Altså har vi i følge testen oppfylt nok et krav ved minste kvadrats metode.

8.3.3 Normalitet

Om feilleddet er normal fordelt sjekkes ved å kjøre Jarque-Berra test i EViews.



Figur 8.3.3

Testen gir p-verdi på 0,557, noe som er høyere enn forkastningsnivået på 5 %. Dermed kan ikke nullhypotesen om at feilleddene er normalfordelte forkastes. Testen viser at verdien for kurtosis er høy, men ikke høy nok til at nullhypotesen om normalfordeling forkastes. Jeg vil også her se om modellen får vesentlige bedre resultater ved å fjerne de boligene som gir størst avvik. De to verdiene lengst til høyre og en til venstre, som er bolig nr. 55, 66 og 166 i datautvalget.

Den nye regresjonen uten disse tre observasjonene ga ingen forbedring av modellen. De samme variablene holdt seg signifikante og forklaringsgraden ble så å si det samme. Det viser seg nok engang at det ikke er noe å hente på å fjerne de største avvikene så lenge feilleddene er normalfordelte. Konklusjonen blir selvsagt å benytte alle observasjonene.

8.3.4 Multikolinearitet

	Sov	Arna	Bad	Åsa.	Balk.	Berg	Bredb.	BTA	Eie.ty.	Gjeld	Fliser	Gar.	Hage	K.tv	Kj.sen	Inla	oppu.	Parket	Sent.	Tomt	V.k.
Sov	1																				
Arna	0,33	1																			
Bad	0,17	0,12	1																		
Åsane	0,3	-0,39	0,03	1																	
Balk.	-0,45	-0,19	-0,13	-0,21	1																
Berg	-0,57	-0,5	-0,12	-0,6	0,36	1															
Bred	-0,35	-0,33	-0,15	-0,08	-0,26	-0,36	1														
BTA	0,35	0,4	0,2	-0,01	-0,3	-0,34	-0,29	1													
Eie.ty.	0,55	0,34	0,17	0,35	-0,68	-0,63	-0,33	0,44	1												
Gjeld	-0,09	0	-0,02	-0,11	0,08	0,13	0,16	-0,14	-0,2	1											
Fliser	-0,18	-0,26	-0,1	-0,06	-0,16	-0,29	-0,29	-0,24	-0,2	0,07	1										
Garas.	0,32	0,01	0,04	0,44	-0,33	-0,42	-0,14	0,19	0,48	-0,08	-0,03	1									
Hage	0,5	0,36	0,15	0,4	-0,54	-0,67	-0,38	0,44	0,82	-0,17	-0,32	0,47	1								
K.tv	-0,48	-0,53	-0,21	-0,15	0,3	0,6	0,6	-0,43	-0,4	0,08	0,21	-0,27	-0,48	1							
Kj.sen	0,41	0,84	0,08	-0,11	-0,31	-0,64	-0,32	0,5	0,52	-0,06	-0,2	0,19	0,5	-0,53	1						
Inald	0,07	0,04	0,02	-0,22	-0,04	0,17	-0,15	0,16	0,01	-0,33	-0,13	0,13	0	-0,11	-0,03	1					
Oppu	-0,03	-0,05	0,02	-0,13	0,11	0,17	-0,04	-0,05	-0,1	-0,24	0,32	-0,14	-0,09	-0,05	-0,13	0,53	1				
Park	-0,21	-0,33	0,05	0,04	-0,03	0,25	0,3	-0,22	-0,1	0,06	0,44	-0,01	-0,2	0,3	-0,26	-0,1	0,15	1			
Sent.	0,51	0,8	0,06	0,17	-0,36	-0,85	-0,36	0,47	0,6	-0,09	-0,26	0,3	0,6	-0,6	0,93	-0,1	-0,16	-0,29	1		
Tomt	0,42	0,46	0,33	0,23	-0,41	0,62	-0,36	0,53	0,61	-0,18	-0,3	0,34	0,67	-0,55	0,53	0	-0,16	-0,27	0,59	1	
V.k	-0,09	-0,16	0	-0,09	0,12	0,22	0,26	-0,24	-0,2	0,09	0,81	-0,02	-0,3	0,18	-0,13	-0,2	0,3	0,51	-0,2	-0,3	1

Tabell 8.3.4 - De største korrelasjonene er merket gult. Resten er merket rødt for negativ og blått for positiv korrelasjon.

Det ser ikke ut som det er store problemer med multikolinearitet her heller. Korrelasjonene mellom variablene som ble brukt i første modell er de samme da dataene ikke er blitt forandret. De største korrelasjonene var mellom ”avstand til sentrum” og ”avstand til nærmeste kjøpesenter” på 0,92 og mellom hage og eiendom på 0,82. Den nye korrelasjonsmatrisen inneholder de tre nye stedsvariablene. Korrelasjonen mellom Arna og ”nærmeste kjøpesenter” med verdien er 0,84 er høy. Dette er som forventet da boligene fra Arna er de som er lengst unna kjøpesenter. Variabelen kan ikke fjernes da det vil ekskludere modellen fra å brukes i Arna området. Det samme gjelder for variablene Bergen og sentrum hvor verdien er -0,85. Negativt fortegn betyr at de samsvarer motsatt. Egentlig skulle man tenke seg at verdien burde vært -1 da boliger i sentrum har fått avstanden 0 km til sentrum. Grunnen til at dette ikke stemmer er at boliger like utenfor sentrum har blitt plassert som bolig i Bergen sentrum og de har ikke verdi 0 km for avstand til sentrum.

Jeg kjørte regresjonen på nytt uten variabelen ”avstand til nærmeste kjøpesenter”, men det ga ikke bedre modell. Velger da som med den andre modellen å beholde variabelen.

8.4 Forbedring av den nye modellen

Samme framgangsmåte som jeg benyttet ved analysen av den første modellen blir brukt. Jeg kjører regresjonen med kun de variablene som ble signifikante først, og sjekker for heteroskedastisitet og autokorrelasjon. Resultatet sees i utskriftene nedenfor.

Dependent Variable: LOGSALGSPRIS_PRM2
Method: Least Squares
Date: 11/27/05 Time: 21:56
Sample: 1 200
Included observations: 200

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ANT_BAD	-0.051808	0.020811	-2.489500	0.0137
ANT_SOV	-0.079850	0.016167	-4.939165	0.0000
ARNA	4.318107	0.062581	69.00011	0.0000
ASANE	4.380722	0.046646	93.91505	0.0000
BERGEN	4.489860	0.018828	238.4705	0.0000
BREDBAND	0.031333	0.015191	2.062614	0.0405
BTA	-0.000455	0.000237	-1.921388	0.0492
EIENDOMSTYPE	0.062442	0.020914	2.985676	0.0032
FE_GJELD	-2.54E-07	4.43E-08	-5.737922	0.0000
HAGE	-0.054070	0.021013	-2.573176	0.0109
KABEL_TV	0.036268	0.017299	2.096489	0.0374
KJ_SENTR	0.010791	0.004602	2.345112	0.0201
SENTRUM	-0.013657	0.004878	-2.799546	0.0057

R-squared	0.862612	Mean dependent var	4.284549
Adjusted R-squared	0.853795	S.D. dependent var	0.203362
S.E. of regression	0.077759	Akaike info criterion	-2.207614
Sum squared resid	1.130688	Schwarz criterion	-1.993223
Log likelihood	233.7614	Durbin-Watson stat	1.938578

Tabell 8.4.1

White Heteroskedasticity Test:

F-statistic	1.233024	Probability	0.246478
Obs*R-squared	19.46287	Probability	0.245392

Tabell 8.4.2

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.050266	Probability	0.998432
Obs*R-squared	0.275804	Probability	0.998074

Tabell 8.4.3

Alle verdiene holder seg signifikante (tabell 8.4.1), samtidig som modellen ikke har problemer med heteroskedastisitet (tabell 8.4.2) og autokorrelasjon (tabell 8.4.3). Tabellene for heteroskedastisitet og autokorrelasjon viser at p-verdien er enn 0,05 som betyr at antagelsene om feilledd med konstant varians og korrelasjon lik null er oppfylt.

Oppgaven videre var da som for første modell å se om noen av de andre variablene ville bli signifikante sammen med variablene i tabell 8.4.1, og om de løse problemet med heteroskedastisitet. Jeg valgte å kjøre regresjonen med variablene fra tabell 6.5.1 mot de resterende variablene hver for seg for å se om noen av variablene ville bli signifikante. Jeg begynte med de med høyest t-verdi først. Etter hver regresjon sjekket hvordan det gikk med heteroskedastisitet og autokorrelasjon. Resultatet er ført inn i tabell 8.4.4.

Variabel	Justert R2	T-verdi	Prob	Prob heteros.	Prob auto.
De signifikante	0,854	>2	<0,05	0,55	0,99
Lnalder	0,855	-1,43	0,153	0,344	0,99
Tomt	0,855	1,56	0,118	0,359	0,96
Balkong	0,854	1,34	0,18	0,354	0,99
Oppusning	0,853	0,03	0,97	0,181	0,99
Fliser	0,853	-0,34	0,72	0,04	0,99
Parkett	0,853	-0,34	0,73	0,138	0,99
Varmekabler	0,853	-0,05	0,96	0,06	0,99
Garasje	0,853	-0,15	0,88	0,28	0,99

Tabell 8.4.4

Ingen av variablene ble signifikante. Selv om det ikke ble noen problemer med heteroskedastisitet og autokorrelasjon er det uinteressant da de ikke er signifikante. Det kunne være fristende å ta med Lnalder, tomt og balkong da de relativt sett ikke er langt fra å være signifikante. Men ut fra statistisk teori skal jeg kun benytte meg av de variablene som blir signifikante. Dessuten ser jeg at de har liten forklaringskraft. Alle tre variablene gir en økning i den justerte forklaringsgraden med 0,001, som er nesten ingenting.
Den endelige modellen blir da å kjøre med de 13 variablene som ble signifikante.
Regresjonen blir:

$LogPm^2$ = -0.05180796252*BAD - 0.0798498385*SOV + 4.31810726*ARNA + 4.380722439*ASANE + 4.489859886*BERGEN + 0.03133344013*BREDBAND - 0.0004549778729*BTA + 0.06244156548*EIENDOMSTYPE - 2.539373219e-007*GJELD - 0.05407049817*HAGE + 0.03626767772*KABEL_TV + 0.01079104931*KJ_SENTR - 0.01365689369*SENTRUM

8.5 Simulering

De samme 20 observasjonene som jeg benyttet i den første modellen blir brukt for å kunne sammenligne modellene mot hverandre. Resultatene er utarbeidet ved hjelp av Excel og ført inn i tabell 8.5 nedenfor.

Nr	Salgspris	Prisantydning	Prisantydning fra megler	% forskjell
1	1745028	1584199	1450000	9,26 %
2	1027466	923458	990000	-6,72 %
3	995471	893997	990000	-9,70 %
4	2273110	2070464	2290000	-9,59 %
5	2018692	1836192	2000000	-8,19 %
6	3450599	3154712	2990000	5,51 %
7	877473	785343	890000	-11,76 %
8	587893	518693	490000	5,86 %
9	1867496	1696944	1650000	2,85 %
10	1170892	1055527	980000	7,71 %
11	1223763	1104211	1190000	-7,21 %
12	1950251	1773171	1800000	-1,49 %
13	1577126	1429592	1390000	2,85 %
14	1560968	1414714	1590000	-11,02 %
15	2061908	1875987	1690000	11,01 %
16	2441718	2225720	2200000	1,17 %
17	1287267	1162687	1350000	-13,88 %
18	1363059	1232477	1490000	-17,28 %
19	1568540	1421686	1390000	2,28 %
20	1915733	1741386	1890000	-7,86 %
	Gjennomsnitt	1495058	1535000	-2,81 %
	Standard avvik	596220	571089	8,46 %

Tabell 8.5

Resultatene er ved første øyekast meget oppløftende. Gjennomsnittelig prosentvis avvik er på -2,81 %. Det vil si at modellen i gjennomsnitt gir 2.81 % lavere prisantydning enn meglerne. Dessuten er avvikene ikke avskrekkende store vist ved standardavvik på 8,46 % selv om avvikene strekker seg fra 11 % til -17,28 %. Sammenlignet med første modell er resultatene vesentlig bedre. Også her er gjennomsnittet og standardavviket verdier av de prosentvise forskjellene mellom boligene og ikke forskjellen mellom totale priser. Det blir benyttet i testen nedenfor.

Jeg utfører den samme ”Two-sample T-test” som ved første modell. Skriver ikke opp hypotesene og formlene på nytt da de er de samme som på side 37. Utregning gir t-verdi på minus 0,216 som betyr at nullhypotesen ikke forkastes da kritisk verdi som sagt er ± 2,021.

Ved hypotese testing på 5 % nivå kan man ikke si at det er forskjell mellom prisantydningen gitt ved modellen og meglerne. T-verdien her er mindre enn ved den første modellen som er et godt tegn. Det vil si at sjansen for at type-2 feil er mindre her. Type-2 feil betyr sannsynligheten for å godta nullhypotesen selv om den egentlig skulle vært forkastet (Lillestøl 1994).

Konklusjonen er at modellen gir et bra estimat på prisantydningen. Modellen oppfyller alle forutsetningene ved minste kvadrats metode samtidig som testutvalget (de 20 observasjonene) består hypotese testing.

9. Kommentarer til regresjonsmodell 2

Jeg vil kort kommentere hva de enkelte variablene betyr, hva de tilfører modellen og hvorfor noen av variablene fikk forskjellig resultat i forhold til hypotesene jeg satte opp. Et minus ved modellen er at det er vanskelig å sette et bestemt beløp på hva de enkelte variablene tilfører da den avhengige variabelen er på logaritme form med 10 som grunntall. Koeffisientene sier altså hvor mye logaritmen til salgsprisen pr kvadratmeter øker eller synker. Jeg forsøker likevel å gi et tall på hver variabel ved å se hva variabelen tilfører konstantleddet når de andre variablene settes lik null.

Arna:
Koeffisienten til Arna er 4,3181076 og har den samme betydningen som et konstantledd. Gitt at alle de andre variablene blir null gir dette en boligpris på 20 802 kroner pr kvadratmeter. Dette virker som et fornuftig estimat selv om gjennomsnittelig pris for de 50 boligene fra Arna ble 11 366 (4.05565). Prisen vil bli trukket mye ned på grunn av lang avstand til sentrum. T-verdien på 69 sier at variabelen er signifikant på 5 % med god margin.

Åsane:
Koeffisienten til Åsane er 4,38072 som gitt alle andre variabler lik null en boligpris pr kvadratmeter på 24 028. Gjennomsnittelig pris for de 63 boligene fra Åsana ble 15 721 (4,19649). Det stemmer bra da at konstantleddet for Åsane er større enn Arna. T-verdien ble 93,9 og er signifikant forskjellig fra null.

Bergen sentrum:
Koeffisienten til Bergen sentrum er 4,49896 og gir boligpris pr kvadratmeter på 30 893 gitt alle andre variabler lik null. Gjennomsnittet for de 87 boligene fra Bergen sentrum ble 30 190 (4,47986) som er omtrent det samme som konstantleddet. Dette stemmer bra da avstandsvariablene blir null. T-verdien ble 238 og er signifikant.

Bad:
Som forventet ble koeffisienten for bad negativ med verdien -0,0518079. Hypotesen om at boligpris pr kvadratmeter faller med ekstra bad stemmer. Med utgangspunkt i konstantleddet for de tre plassene og at andre variabler er lik null, reduserer ekstra bad prisen med 2 339 for

Arna, 2 701 for Åsane og 3 474 for Bergen sentrum. T-verdien ble -2,49 og signifikant med p-verdi 0,0137.

Soverom:
Hypotesen ble som forventet at prisen pr kvadratmeter faller der det er mer enn et soverom med koeffisient på -0,07985. Gitt konstantleddet og at de andre variablene er lik null reduserer to soverom eller mer prisen med 3 493 for Arna, 4 035 for Åsane og 5 188 for Bergen sentrum. T-verdien ble -4,94 med p-verdi på 0,000.

Bredbånd:
Hypotesen om at bredbånd øker prisen pr kvadratmeter stemmer med positiv koeffisient på 0,031333. Ved å sette alle variablene lik null bortsett fra konstantleddet øker prisen med 1 556 for Arna, 1 798 for Åsane og 2 311 for Bergen sentrum. Det virker unaturlig at prisen øker med så mye fordi det er bredbånd innlagt. For Arna man forvente at innlagt bredbånd vil ha stor verdi da det antakelig ikke er muligheter for å få innlagt bredbånd i de mer fjernliggende områdene enda. Men for Åsane og Bergen virker dette veldig unaturlig da tilgangen på bredbånd forventes å være større der. Grunnen til disse høye tallene er at modellen opererer med 10 som grunntall slik at vi har eksponsiell vekst her. En økning fra 4,4 til 4,5 med 10 som grunntall tilsvarer en økning fra 25 118 til 31 622 med ”normale” tall. Poenget er at man må se sammenhengen mellom alle variablene.
T-verdien ble 2,06 med en p-verdi på 0,0405.

BTA:
Hypotesen om at prisen pr kvadratmeter skulle øke med antall kvadratmeter tilleggsareal stemte ikke ifølge regresjonen. Koeffisienten ble negativ med verdi -0,000455. Grunnen til dette resultatet kan nok skyldes at tilleggsarealet samsvarer med boarealet. Med det mener jeg at der det er høyt boareal er det mest sannsynlig høyt tilleggsareal. Som sagt forventes boligprisen pr kvadratmeter å være høyere for mindre boligenheter slik at tilleggsarealet trekker ned boligprisen pr kvadratmeter. Gjennomsnittelig var tilleggsarealet på 25 m^2. Gitt alle variabler bortsett fra konstantleddet lik null gir 25 m^2 tilleggsarealet en reduksjon i prisen på 537 for Arna, 620 for Åsane og 789 for Bergen sentrum. Men må ikke legge så mye i disse tallene da BTA varierer fra 0 til 275 m^2 for de 200 observasjonene.
T-verdi ble -1,92 med p-verdi på 0,0492 og signifikant. Av de variablene som ble signifikante var BTA den som var nærmest fra å bli forkastet.

Eiendomstype:

Hypotesen om at eneboliger skulle redusere salgsprisen pr kvadratmeter ble heller ikke som forventet. Koeffisienten ble positiv med verdi 0,062442. Grunnen til dette har jeg ingen god forklaring på. Riktig nok kunne man tenke seg at eiendomstype og hage som hadde høy korrelasjon på 0,82 ville kunne påvirke hverandre. Siden de ikke forandret fortegn eller størrelse av betydning når jeg kjørte regresjonen uten den andre tilstede skulle det ikke tyde på problemer med multikolinearitet. Gitt alle variablene lik null bortsett fra konstantleddet, øker enebolig prisen pr kvadratmeter med 3 216 for Arna, 3 716 for Åsane og 4 777 for Bergen sentrum. T-verdi ble 2,96 som gir p-verdi 0,0032 og signifikant variabel.

Gjeld:

Koeffisienten for gjeld ble -0,0000002539 som stemte bra med hypotesen om at gjeld trekker ned salgsprisen pr m^2. Ikke så mye mer å kommentere enn at logaritmen til salgsprisen pr kvadratmeter synker med -0,0000002539 pr krone gjeld. I den forrige modellen var det lettere å kommentere da man enkelt kunne se hvor stor andel av gjelden som påvirket salgsprisen. Variabelen er signifikant med t-verdi -5,74 og p-verdi 0,000.

Hage:

Koeffisienten for hage ble -0,05407 som betyr at prisen faller med et tillegg der det er hage. Resultatet er det motsatte av hypotesen jeg satte opp. Grunnen til det kan kanskje skyldes at hage ofte er tilstede der hvor boarealet er stort noe datautvalget støtter oppunder. Stort boareal trekker prisen pr kvadratmeter ned slik at hage da har fått negativ verdi. Jeg noterer meg at hage og eiendomstype som hadde korrelasjonskoeffisient på 0,82 nesten utvisker hverandre hvis begge er tilstede (0,062442-0,05407). Ved å sette variablene lik null og legge til konstantleddet reduserer tilstedeværelsen av hage prisen med 2 435 for Arna, 2 812 for Åsane og 3 616 for Bergen sentrum. T-verdien ble -2,57 med p-verdi 0,01.

Kabel-TV:

Hypotesen om at salgsprisen pr kvadratmeter får et tillegg for kabel-tv stemmer med resultatet fra modellen hvor koeffisienten ble 0,036267. Ved å ta utgangspunkt i alle variabler lik null bortsett fra konstantleddet øker prisen pr kvadratmeter der det er kabel-tv med 1 811 for Arna, 2 093 for Åsane og 2 690 for Bergen sentrum. Jeg gir den samme kommentaren som for bredbånd at det virker unaturlig at det å ha kable-tv øker prisen så mye, men at det har en sammenheng med resten av modellen. T-verdien ble 2,1 med p-verdi 0,037.

Avstand til sentrum:

Koeffisienten ble -0,013657 som henger i hop med hypotesen om at prisen pr kvadratmeter synker med avstanden fra sentrum. Boliger i sentrum får null fratrekk mens boliger i Arna har størst nedtrekk. T-verdien ble -2,8 med p-verdi 0,0057.

Avstand til nærmeste kjøpesenter:

Hypotesen om at prisen synker med avstand til nærmeste kjøpesenter stemte ikke med modellen. Koeffisienten ble 0,01079 slik at prisen øker med avstand fra kjøpesenteret. Resultatet virker ulogisk, men har prøvd å gi en forklaring på dette på side 33. En bolig på Arna vil få økt verdi i prisen på grunn av avstanden til kjøpesenter, men ser at virkningen av avstanden til sentrum er større slik at sum summarum blir prisen redusert. T-verdien ble 2,35 med p-verdi 0,02.

10. Prisantydning vs. Salgspris og verditakst

Som en tilleggsoppgave ønsket jeg å se på hvor stort avviket mellom prisantydningen og salgsprisen var. I en artikkel jeg fant på Forbrukerombudet sine hjemmesider fra april i år (www.forbrukerombudet.no), kom det fram at det ved stikkprøver i de største byene i Norge var over 20 % av boligene som ble solgt med en pris som var 10 % høyere enn prisantydningen. Verst ut kom Bergen hvor hele 45 % av boligene hadde en salgspris som var mer enn 10 % høyere enn prisantydningen. Forbrukerombudet går langt i å påstå at dette er en bevist prising fra bransjen for å lokke til seg kjøpere da boligen virker å være "billig". Det antydes altså at meglerne bevist priser boligen lavere enn meglers objektive prisantydning ville vært. Utsagnet forsvares med at en rekke boliger får en prisantydning som er lavere enn det selgeren av boligen er villig til å selge det for. Videre har de sett på hvilke eiendomsselskap som avviker mest, og for Bergen er det Aktiv eiendomsmegling og Garanti eiendomsmegling som avviker mest.

Ved å bruke datautvalget mitt på 200 boliger fra Bergens området sjekket jeg om mine data stemte overens med stikkprøvene foretatt av Forbrukerombudet. Ved hjelp av Excel kom jeg fram til at 49 % av de 200 boligene i datasettet hadde en salgspris som var 10 % høyere enn prisantydningen. Dermed fikk jeg nesten samme resultat som Forbrukerombudet som hadde 45 %. Gjennomsnittet av datautvalget ga en salgspris som lå 10,1 % høyere enn prisantydningen som tyder sterkt på en bevist underprising av boligene. I 89 % av tilfellene var salgsprisen høyere enn prisantydningen.

Videre så jeg på om det var noen av eiendomsselskapene som skilte seg mer ut enn andre. Resultatene er satt inn i tabell 10 nedenfor.

Selskap	Antall	Antall over 10 %	Antall over 0 %	Gjennomsnitt
Aktiv eiendomsmegling	46	52,2 %	100,0 %	10,3 %
DnbNor eiendom	75	56 %	90,7 %	9,5 %
Eiendomsmegler Vest	55	47,3 %	94,5 %	11,5 %
Garanti eiendomsmegling	4	50 %	100 %	11,8 %
Øvrige selskap	20	20 %	85 %	8,2 %

Tabell 10

Ved å bruke datautvalget mitt blir det DnbNor som får flest boliger hvor salgsprisen er 10 % høyere enn prisantydningen med en andel på 56 %. Ved å se på gjennomsnittsprisen er det

Garanti eiendomsmegling som har størst avvik med salgspris 11,8 % høyere enn prisantydning. Men med bare 4 observasjoner må enn være forsiktig med å konkludere noe. De fire selskapene som er nevnt her skiller seg veldig lite fra hverandre. I alle fall for lite til å konkludere at det ene selskapet har større avvik enn de andre. Det man kan si er at de øvrige selskapene som er fordelt på 10 selskap skiller seg ut fra de fire. Her har kun 20 % av boligene enn salgspris som er 10 % høyere enn prisantydningen. Dessuten har de det minste gjennomsnittelig avviket med en salgspris på 8,2 % høyere enn prisantydningen.

Konklusjonen min blir den samme som for Forbrukerombudet at meglerne bevist ser ut til å sette prisantydningen lavere enn den burde vært. Derimot kan jeg ikke si at noen av selskapene er verre enn de andre.

11. Konklusjon

Konklusjonen i min oppgave er at modellen jeg har kommet fram til fungerer bra for området Arna, Åsane og Bergen sentrum. Den oppfyller alle antagelsene for minste kvadraters metode som er en forutsetning for en god modell. Det har blitt stilt strenge krav til modellen med tanke på signifikante verdier. Hypoteser som ikke ble som forventet har derimot blitt godtatt da de har vært signifikante. Modellen ble brukt til å simulere prisantydningen som stemte bra med prisantydningen satt av eiendomsmeglere. ”Two sample” testen konkluderte med at det ikke var forskjell på modellens og meglernes prisantydning. Tilleggsoppgaven tilslutt viste at mine data stemte bra med data brukt av Forbrukerombudet da vi kom fram til samme konklusjon. Det gir meg en trygghet på at jeg har funnet et datautvalg som er nokså identisk med området modellen min gjelder for.

Modellen jeg har kommet fram til er enkel og rask å bruke samtidig som den gir et tilfredsstillende estimat av boligprisen. De opplysningene som trenges er å finne på de aller fleste salgsprospekter. Modellen er ment å kunne brukes til å gi et raskt og enkelt estimat på boligprisen både for selger og kjøper.

Litteraturliste

Brooks, Chris. Introductory Econometrics for Finance, The ISMA Centre, University of Reading, 5th edition 2004

Gule Sider, Hordaland 2005

Jensen, Christine 2005. Bygg balkong – få 600 00 ekstra. Verdens Gang. <http://www.vg.no/pub/vgart.hbs?artid=296025> (8. nov 2005)

Keller, Gerald & Warrack, Brian Statistics for Mangement and Economics, 6th edition 2003

Lillestøl, Jostein. Sannsynlighetsregning og statistikk med anvendelser, Cappelen Akademiske Forlag, 1994

Norges Eiendomsmeglerforbund (NEF), www.nef.no

Norsk Standard, NS 3940

Samtale med meglere: Knut Meeg Torbjørnsen, Notar eiendom. Magnus Dehli, Pareto eiendom.

Rosen, Sherwin "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", Journal of Political Economy Jan 1974, no 82

www.eiendomsverdi.no

www.forbrukerombudet.no, <http://www.forbrukerombudet.no/index.gan?id=11002120&subid=0>(6. April 2005)

www.finn.no

www.kvasir.no

	Arna	Åsane	Bergen	Alder	Bad	Soverom	Balkong	BOA	Bredb	BTA	Eiend.	Gjeld	Fliser	Garasje	Hage	Kabel-tv	Kj.sent	Sentrum	Tomt	Varmek	Salgspris
Mean	0.25	0,315	0.435	54,3	0.1	0,725	0,285	99,115	0,035	24,785	0,54	39517	0,815	0,29	0,53	0,52	5,03	5,62	321,69	0,84	1786460
Median	0	0	0	41	0	1	0	100	0	17	1	0	1	0	1	1	2,7	10,1	111,5	1	1705000
Maximum	1	1	1	168	1	1	1	248	1	275	1	900000	1	1	1	1	24,5	33,3	1758	1	4100000
Minimum	0	0	0	1	0	0	0	21	0	-97	0	0	0	0	0	0	0	0	0	0	620000
Std. Dev.	0,434	0,465	0,497	41,88	0,3	0,44	0,453	47,88	0,462	29,47	0,5	128484	0,389	0,455	0,5	0,5	5,91	9,18	421,65	0,367	637456
Skewness	1,155	0,797	0,262	0,76	2,66	-1,008	0,953	0,628	0,847	3,45	-0,16	3,88	-1,62	0,923	-0,12	-0,08	1,41	0,55	1,345	-1,85	1
Kurtosis	2,333	1,634	1,068	2,48	8,11	2,012	1,907	2,909	0,717	29,77	1,025	19,16	3,632	1,856	1,01	1	4,02	2,2	3,9	4,44	4,39
Jarque-Bera	48,15	0,69	33,372	21,55	455	41,93	40,193	13,164	37,623	6371	33,33	2681	91,08	39,44	33,3	33,33	74,71	15,45	67,22	131,97	49,71
Probability	0	0	0	0	0	0	0	0,001	0	0	0	0	0	0	0	0	0	0	0	0	0
Sum	50	63	87	10860	20	145	57	19823	61	4957	108	7903586	163	58	106	104	1006	1924	64337	168	3.57E+08
Sum Sq. Dev.	37,5	43,155	49,155	349180	18	39,875	40,755	456228	42,395	172863	49,68	3,29E+12	30,16	41,18	49,8	49,92	6958	16772	3,54E+07	26,88	8.09E+13
Observations	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200

Vedlegg 1- beskrivende data