Stockholms universitet

# Att förutse urbaniseringsnivå i Sveriges kommuner med betaregression

Oliver Ng

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2019:28
Matematisk statistik
Augusti 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2019:28**
http://www.math.su.se/matstat

# Att förutse urbaniseringsnivå i Sveriges kommuner med betaregression

Oliver Ng*

Augusti 2019

**Sammanfattning**

Syftet med denna uppsats är att modellera och analysera vilka faktorer som förklarar urbaniseringsgraden i Sveriges kommuner. Regressionsanalys med betaregression används för att skapa en modell med signifikanta faktorer. Data är baserad på två perioder, 2010 och 2015, med 8 olika parametrar för alla kommuner i Sverige. Med hjälp av variabelselektion visas att två faktorer, nuvarande urbaniseringsgrad och tätortarea, var det största bidragande faktorerna för urbaniseringsgraden i framtiden. Modellen visar sig även kunna modellera framtida urbanisering till en utmärkt grad men det ifrågasätts om den kan användas under andra omständigheter.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: obvious.spycrab@gmail.com. Handledare: Martin Sköld & Tom Britton.

**Förord**

Jag vill tacka mina handledare, Tom Britton och Martin Sköld för deras handledning och feedback för att förbättra uppsatsen teori och grammatik. Dessutom vill jag tacka R studio som utgjort grunden i analysarbetet, i **R** så var **Betareg** och **tidyverse** högst användbara tillägg.

# Innehåll

# 1 Introduktion

Idag finns ungefär 8 miljarder människor i världen, en del av befolkningen bor utspridd i landskapet när i många länder, en större andel bor i städer eller tätorter. Andelen människor som bor i dessa tätorter kallas urbaniserings- eller tätortsgraden, från lokala byar till globala skalan.

En svensk tätort är definierat enligt Statistiska CentralByrån (SCB) som ett område med befolkningstäthet över 200 $personer/km^2$. Idag bor drygt 80 % av Sveriges befolkning i tätorter men dess urbaniseringsgrad skiljer sig på kommunnivå, syftet med uppsatsen är att undersöka vilka faktorer som förklarar urbaniseringsgraden i kommunerna med hjälp av regressionsanalys.[2]

Vi har inte hittat någon liknade studie med regressionsanalys för tillväxt, det närmaste som fanns var en studie[7] som undersöker städers förutsättning för tillväxt. I denna studie hänvisade de till faktorer som vilken typ av stad, transportering, näringsliv och utbildningsnivå som utgör grunden för tillväxt i framtiden.

## 1.1 Syfte och metod

Denna uppsats syftar till att besvara följande frågor:

- Kan en modell på en tillfredställande sätt prediktera framtida urbaniseringsgrad på kommunnivå?
- Vilka förklarande variabler har modellen och vad är dess styrka?
- Vad är den minsta möjliga modell av signifikanta variabler?
- Vilka implikationer har modellerna?

För att följa upp på dessa frågor används regressionsanalys för att skapa en modell av förklarande variabler. Med hjälp av variabeltransformation och -selektion förenklas modellen och efteråt testas modellens prediktionsförmåga. Eftersom responsen kommer vara procentuell används Betaregression för nästan all modellering.

Majoritet av analysen genomförs i programmet **RStudio** med tillägget **Betareg** för att utföra betaregression. Betaregression kommer att förklaras mer genomförligt i Avsnitt 2 med andra övriga aktuella teorier i uppsatsen.

Avsnitt 3 analyseras data till modellen och när möjligt förenklar och redigerar om data till en annan format. Dessutom tas några kommuner bort från datasetet som sen används för en prediktions test senare i uppsatsen.

Avsnitt 4 inleds regressionsanalys, med både beta- och linjär regression.

Avsnitt 5 är slutsats och tolkning av modellerna som är signifikanta och dess implikation.

Avsnitt 6 är fortsatt diskussion om eventuella brister i modellen och förslag för ytterligare arbete.

# 2 Teorier och metoder.

I detta avsnitt går vi igenom teori och metoder som uppsatsen applicerar.

## 2.1 Betaregression

Betaregression är en variant av regression, den används när man önskar att modellera proportioner, sådant som bensinhalt eller andra intressanta intervalls värde.

### 2.1.1 Varför Betaregression

Första hand, eftersom vi önskar att modellera proportioner som exempelvis urbaniseringsgraden rekommenderas inte linjär regression då modellen kan anta värdet utanför enhetsintervallet. Vilket är inte en rimligt utfall för den verkliga modellen.

Det vore möjligt att parameterisera om data till en länk-funktion som löser föregående problem, men det skulle fortfarande innebära andra komplikationer som att modellen blir heteroskedastisk, assymmetriskt. Därför med hjälp av beta tolkningen som motverkar heteroskedastisk och en länk-funktion gör det möjligt att modellera inom en intervall. [6]

### 2.1.2 Definition

Betaregression är baserad runt antagelse att responsen i datasetet följer en beta fördelning, definierat nominellt:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1},\ 0 < y < 1 \tag{1}$$

Låt oss nu parametrarisera om fördelningen, så att $\mu = p/(p+q)$ och $\phi = p+q$, det följer:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \sim B(\mu,\phi), 0 < y < 1 \tag{2}$$

Vi noterar att $\phi$ kan bli tolkat som spridningsmått, då med fixt $\mu$ variansen blir mindre beroende på hur stor $\phi$ är. Ty variansen för betafördelning följer:

$$var(y) = \frac{\mu(1-\mu))}{1+\phi} \tag{3}$$

Låt $y_1, ..., y_n$ vara observationer av oberoende slump variabler, där $y_i, i = 1, ..., n$; är fördelat $\sim B(\mu_i, \phi)$ med känd $\mu$ och okänd spridningsmått $\phi$. Modellen kan då beskrivas som:

$$g(\mu_i) = \alpha + x_i^T \beta = \eta_i, \tag{4}$$
$$\eta_i = \alpha + \beta_1 x_{i1} + ... + \beta_k x_{ik} \tag{5}$$

Där $\beta = (\beta_1, ..., \beta_k)^T$ är en vektor av okända regressions parametrar och $x_i = (x_{i1}, ..., x_{ik}), i = 1, ..., n$ är vektorer av förklarande variabler.

Tekniskt sett är $\alpha$ en $\beta$ där kovariaten är lika med 1, det vill säga ett intercept som i linjär regression.

Sist är $g(.) : (0, 1) \mapsto \mathbb{R}$ en länk funktion som är två gånger deriverbar, i uppsatsen används logitfunktion definierad som[6]:

$$g(\mu_i) = \frac{\mu_i}{1 - \mu_i} \tag{6}$$

### 2.1.3 Skattning

Betaregressions parametervärden skattas numeriskt, eftersom det saknas explicit lösning. Skattning utförs då i **R** ML funktionen **BFGS**, en ML metod för att skatta parmetrarna för modellen. [6]

### 2.1.4 Loglikelihood funktion

Från (2) erhålls även Loglikelihood funktionen, definierad som $l(\beta, \phi) = \sum_{i=1}^{n} l_i(\mu_i, \phi)$ där: [6]

$$l_i(\mu_i, \phi) = log\Gamma(\phi) - log\Gamma(\mu_i\phi) - log\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)logy_i + [(1-\mu_i)\phi - 1]log(1-y_i) \tag{7}$$

## 2.2 Diagnos av modellens anpassning till data

### 2.2.1 Förklaringsgrad

Förklaringsgraden, $R^2$, är ett vanligt mått som används för att översiktligen döma modellens anpassning till data, dock från länkfunktionen som vårat betaregression använder kan vi inte använda den ordinära $R^2$ för linjär regression.

Istället används en liknade mått Pseudo $R^2$, framöver kallat $R^2$, vilket kommer från korrelationskoefficienten emellan länkfunktionen och $\eta$. Värdet ges i $0 \leq R^2 \leq 1$ där 1 betyder exakt match och motsatsen för 0, explicit definierat.

$$Pseudo\ R^2 = corr(g(y), \eta)^2 \tag{8}$$

Lik dess kusin för linjär regression, så gynnas $R^2$ från större modeller även om en del av dess variabler har ingen signifikans för modellen i frågan. [6]

### 2.2.2 Akaike Information Criterion

Akaike Information Criterion(AIC) används för att tolka modellens anpassningsmått, det används för att välja mellan modeller som delar liknande egenskaper såsom responvariabel. Vid sådan jämförelse anses den modellen som har minst AIC värde som lämpligast modell. Det definieras som:

$$AIC = 2k - 2ln(\hat{L}) \tag{9}$$

Där $k$ är antalet parametrar och $\hat{L}$ är skattad likelihood-funktionen för modellen. [5]

Mellan AIC och Förklaringsgraden gynnar AIC istället mindre modeller motsatt till $R^2$. Vilket varför både AIC och $R^2$ tillämpas för att hitta den genomsnittliga bästa modell.

### 2.2.3 P-värde

P-värde av $\beta_k$ är ett sätt att räkna ut hur sannolikt parametern $\beta_k$ är signifikant i modellen. Detta erhålls genom Wald test likt ordinär linjär regression. Låt $\beta$ vara den önskade test parameter, vi vill bestämma p-värdet för hypotesen:

$$H_0 : \beta = \beta_0 \tag{10}$$
$$H_1 : \beta \neq \beta_0 \tag{11}$$

Där $\beta_0$ är oftast lika med 0, vilket ger oss följande test statistika:

$$W = \frac{\hat{\beta_k}^2}{Var(\hat{\beta_k})} \tag{12}$$

P-värdet är sannolikheten att åstadkomma liknade eller extremare resultat till $H_0$ mot det som var observerat, P-värdet erhålls genom:

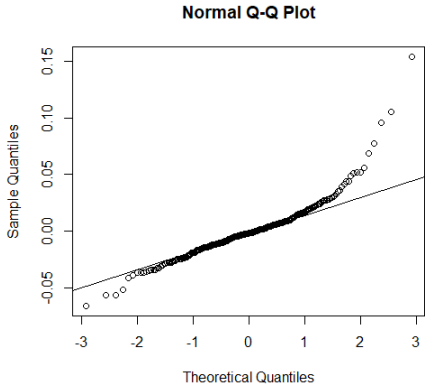
$$P_{\beta_0} = P(|W| > \chi^2(1)_{\alpha/2}) \rightarrow \alpha \tag{13}$$

Där $\chi^2(1)_{\alpha/2}$ är fördelningen för Wald testet. Om vi kan förkasta $H_0$ då P-värdet är mindre än $\alpha$ erhålls då ett hypotestest med signifkansnivå $\alpha$. [5]

### 2.2.4 Residual plot

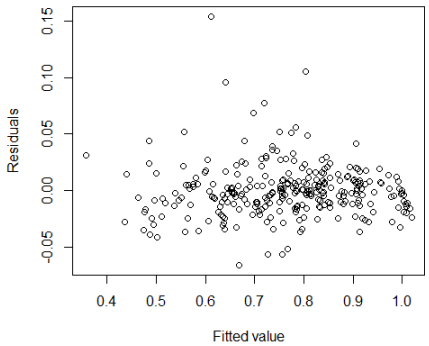
Residualen är skillnaden mellan den skattade värde och det observerat värde.

Ett av kraven för linjär regression säger att residualerna måste vara normalfördelade, annars kan man inte säkert använda modellen. Normal kontroll utför vi med en normal qqplot med exempel från Figur 1




Figur 1: Exempel av qqplot, den prickad linje är förväntad normal för datan som ideallt punkterna ska följa.

Residualernas varians kontrolleras genom att plotta dem mot anpassade värden, om de inte fördelar sig jämnt i plotten tyder det att modellen har inte konstant varians för alla observationer. Alltså ifrågesätts det om modellen har samma fördelning genom hela datasetet, detta tittar vi genom att göra en graf med x-axeln för skattad värde och y-axeln för residualen, som exempel med Figur 2. [4]




Figur 2: Exempel av varians kontroll, ideallt ska alla punkter vara fördelad jämnt över hela grafen

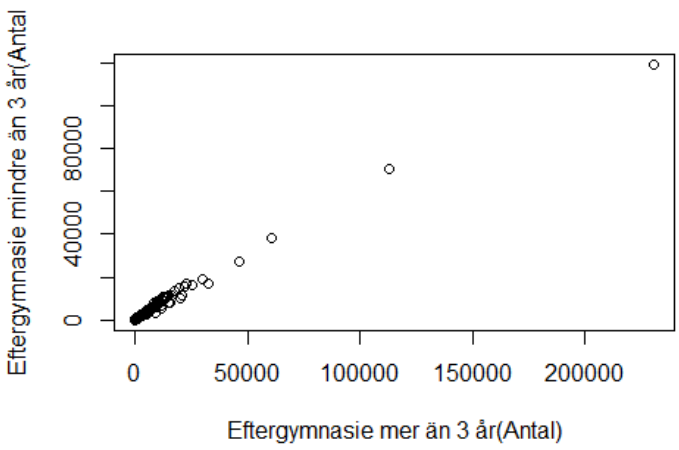
I liknade fall, diagnostiska plottar för Betaregression skiljer sig igen från linjär regression. Residualen kan vara normalfördelat men det är inte ett krav jämfört med linjär regression och det också håller för varians plotten som följer av kraven för normalfördelning, betaregression har däremot andra diagnostiska grafer men är givetvis mer svår tolkat i jämförelse med linjär regression. [6]

## 2.3 Behandling av data

### 2.3.1 Nästan-Kolineära variabler

I många modeller potentiellt finns det förklarande variabler som har linjär samband med varandra, detta implicerar att variablerna avbilda samma faktor. Det kallas då kolineäritet och bör hanteras då modellen får annars onödigt komplexitet.

Dessa kolineäritet kan utredas manuellt, genom att plotta ut två önskade variabler i x och y axel. Om plotten visar en linjär samband så är det stor sannolikhet att det lyder kolineäritet.




Figur 3: Exempel av linjär korrelation, notera linjära sambandet mellan x och y värde

Alternativt används Variance Inflation Factor(VIF), som är ett mått för att finna variablernas kolineäritet, den definieras som:

$$VIF = \frac{1}{1 - R_j^2} \tag{14}$$

Där $R_j^2$ är förklaringsgrad för variation i variabel $x_j$ förklaras av de övriga förklarande variabler. [4]

När en sådan kolinjäritet upptäcks finns det fåtal sätt att bearbeta det, några alternativ är att antingen kombinera kolinjära variabler till en variabler eller så förkastas variabeln då dess likhet finns redan i modellen. [4]

### 2.3.2 Variabel selektion

I regressionsanalys är det inte ovanligt att modeller har icke signifikanta p-värde, det vill säga att variabeln förklarar inte responsvariabeln. Det implicerar att dessa variabler kan då tas bort för att förenkla modellen, vilket kan gynna modellens anpassning till verkliga data.

Det finns tre metoder för välja dessa variabler. *Framåt selektion* då man lägger till variabler i en tom modell, *Bakåt selektion* tas bort icke-signifikanta variabler från modellen och *Stegvis* då man kontrollerar modellen efter varje selektion med framåt eller bakåt selektion. I uppsatsen utnyttjade bara Stegvis selektion då det ansåg onödigt att upprepa processen två ytterligare gånger.

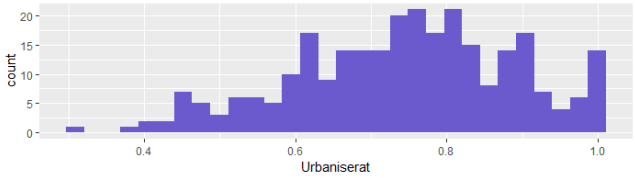
Specifikt utförs selektion på en fullständigt modell genom att exkluderar variablerna med högsta p-värde, återskapar sen modellen och upprepar selektionen tills det inte finns fler icke-signifikanta variabel kvar (om det är mindre än 0.05 i p-värde). I tabellerna framöver kommer det finnas symboler (*) vid p-värdet, det är en indikator av dess p-värde uppdelad i * = 0.05, ** = 0.001, *** = 0. [4]

# 3 Dataanalys

Datamaterialet kommer från Statistiska CentralByrån (SCB), Arbetsförmedlingen (AF) och Sverige Kommuner och Landsting (SKL), det består av 290 kommuner med 8 olika förklarande variabler baserad på observationer från studien[7]. Några variabler är i proportionell format när det resterande är i heltal.[2], [3], [1]

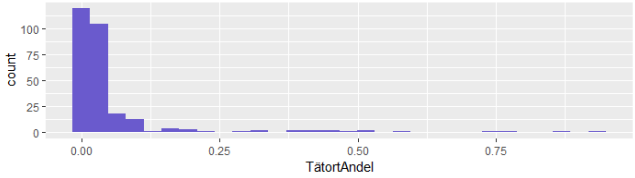
Då vi önskar att modellera urbaniseringen som en proportion vill vi transformera variablerna till enhetsintervallet. Några specifika variabler blev transformerat på annat sätt vilket tas upp i Databehandling, datasetet har 8 följande variabler:

**Urbanisering:**



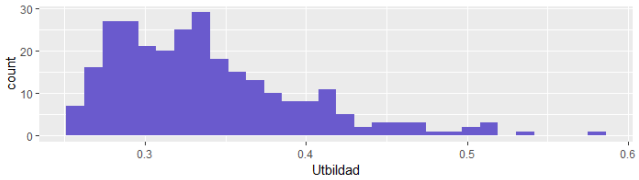

Figur 4: Andel av kommunens befolkning i tätort, består av data från två perioder 2010 och 2015. Graf för År 2010, antar värde mellan [0.31, 1]

**Tätortsarea:**



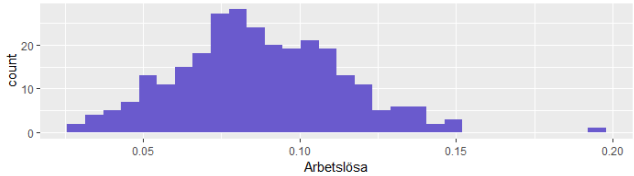

Figur 5: Andel tätort mark i kommunen. Antar värde mellan [0.01, 0.93]

**Utbildning:**



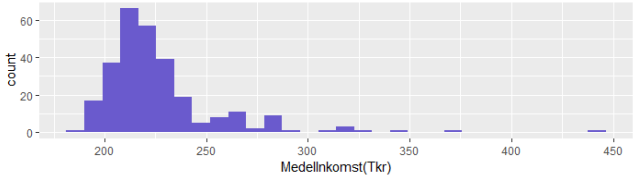

Figur 6: Andel befolkning med utbildning från gymnasiet till forskarutbildning i kommunen, omredigerad till en parameter. Antar värde mellan [0.25, 0.58]

**Arbetslöshet:**




Figur 7: Andel befolkning som är inskrivna som arbetslösa längre än 6 månader hos AF [3]. Antar värde mellan [0.02, 0.19]

**Inkomst:**



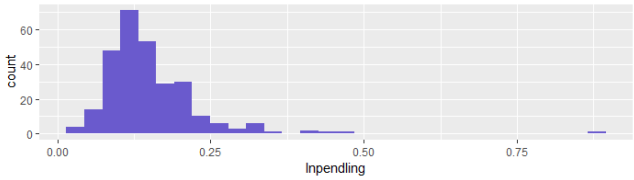

Figur 8: Medelinkomst för kommunens invånare. Antar värde mellan [184, 441]

**Kommunindelning:** :

Tabell 1: Kommunerna enligt SKL [1] kan delas i nio olika kategorier av befolkning som visas i Tabell 1 Kommunindelning

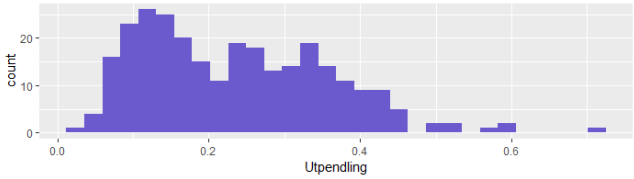
| Kodnr | Beskrivning | Antal |
|---|---|---|
| 1 | Storstäder med mer än 200 000 boende | 3 |
| 2 | Pendling kommun nära storstäder med minst 40 % natt pendling | 43 |
| 3 | Större städer, befolkning mellan 200 000 och 50 000 | 21 |
| 4 | Pendlingskommun nära större städer, minst 40 % natt pendling | 52 |
| 5 | Lågpendling kommuner nära större städer, mindre än 40 % pendling | 70 |
| 6 | Mindre städer, befolkning emellan 40 000 och 15 000 | 29 |
| 7 | Pendlingskommuner nära mindre städer med minst 30 % natt pendling | 52 |
| 8 | Landsbygdkommun med befolkning mindre än 15 000 | 40 |
| 9 | Landsbygkommun med besöksnäring | 15 |

**Inpendling:**




Figur 9: Andel av inpendlade befolkning jämfört med kommunens befolkning. Antar värde mellan [0.03, 0.88]

**Utpendling:**




Figur 10: Andel av utpendling från kommunens befolkning. Antar värde mellan [0.02, 0.71]

## 3.1 Databehandling

I förra sektionen påpekades att förklarande variabler är transformerade, det är genomfört för alla utom en variabel då det ansågs att det inte fanns någon uppenbart sätta att transformera denna. I detta kapitel går vi över beartbetade variabler och reasonemangen bakom transformationerna.

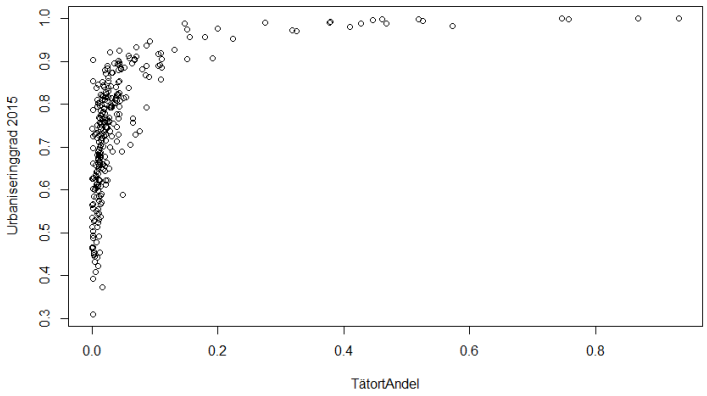
Några exempel på bearbetat transformationer är *Utbildad*, vilket under databehandling visade att faktorn hade högt kolineäritet mellan utbildningsnivåerna. Det visas med VIF i Tabell 2.

Tabell 2: VIF analys för Utbildad

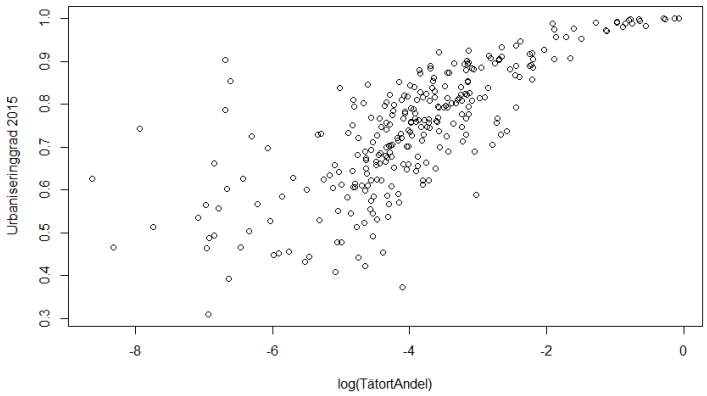
| Variabel | VIF värde |
|---|---|
| Gymnasie | 17.57 |
| Eftergymnasie 3 år | 8.82 |
| Forskarutbildning | 17.53 |

Som tumregel anses det finnas kolineäritet om VIF faktor är större än 3, vilket demonstreras klart i Tabell 2, korrelationen hanterades genom att summera alla variabler till en gemensam variabel som visar andelen befolkning med gymnasie eller högre utbildning *Utbildad*.

En annan intressant korrelation upptäckts mellan *TätortsAndel* och *Urbaniserat*, Figur 11 visar en synligt exponential korrelation. Logaritmerar man *TätortAndel* enligt Figur 12, visas det utspritt men ändå tydligt linjär samband.




Figur 11: Tätort mot Urbanisering

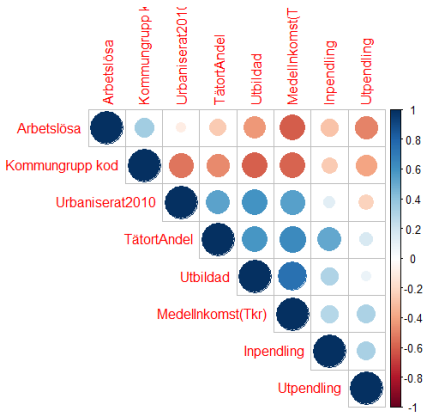



Figur 12: log(Tätort) mot Urbanisering

Ett intressant fall observeras med korrelationen i Figur 12, eftersom urbaniseringsgraden mellan åren är lik varandra implicera att *TätortAndel* är också korrelat med en förklarande och respons variabel. Vi återkommer till detta igen i avsnitt 4.2.

Sist fann vi inga andra intressanta fall i datasetet, därmed blir den slutliga samling i datasetet enligt Tabell 3.

För att även testa precisionen av eventuella modeller, tar vi bort tio slump valda kommuner från datasetet enligt Tabell 4.




Figur 13: Korrelations plot där helt kolineära variabler är 1

Tabell 3: Variablerna för dataset

| Variabel | Beskrivning |
|---|---|
| Urbaniserat2015 | Andel befolkning i tätort år 2015, responsvariabel |
| Urbaniserat2010 | Andel befolkning i tätort år 2010, förklarandevarialer |
| TätortAndel | Andel mark som är tätort bebyggt, |
| Arbetslösa | Andel befolkning som är arbetslösa |
| Kommungrupp | Kategorisk parameter för kommuntyp |
| Utbildad | Andel befolkning med utbildning över gymnasiet |
| Medelinkomst | Medelinkomst av befolkningen i Tusen kr |
| Inpendling | Relativt andel av inpendlade befolkning från andra kommuner |
| Utpendling | Relativt andel av kommunens befolkning pendlar mot en annan kommun |

Tabell 4: Borttagna kommuner

| | |
|---|---|
| Hallsberg | Strömsund |
| Kristinehamn | Vetlanda |
| Munkfors | Luleå |
| Kumla | Växsjö |
| Vindeln | Hässleholm |

# 4 Modellering

Med datasetet färdigt kan modellering utföras, vi börjar kapitlet med att beskriva vilkoren för eventuella modeller.

Frågeställning upprepas, vilket är att hitta den enklaste modellen möjligt för urbaniseringsgraden och förändringen mellan perioderna, reasonemangen hänvisar att med två perioder modellerar vi framtida urbanisering baserad på den föregående period.

För att besvara det, använder vi av responsen *Urbaniserat2015*, med åtta förklarande variabler från Tabell 3. Eftersom responsen är en proportion används betaregression och ordinär linjär regression när det anses dugligt alltså när responsvariabeln är ej en proportion.

Dock att QQ och varians plot är inte krav av betaregression, det kan ändå visa hur dess residual bete sig.

## 4.1 Betamodell

### 4.1.1 Grund modell $M_{Beta}$

Den första modell, $M_{Beta}$ är en enkelt betaregression med alla variabler från Tabell 3, med en logit länk-funktion för $g(.)$ blir dess explicita form för $M_{Beta}$:

$$g(\mu_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} \quad (15)$$

Tabell 5: $M_{Beta}$

| $M_{Beta}$ | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.9256384 | 0.4379751 | -6.680 | 2.39e-11 *** |
| TätortAndel | 6.5921517 | 0.3687549 | 17.877 | 2e-16 *** |
| Urbaniserat2010 | 4.2537127 | 0.1260815 | 33.738 | 2e-16 *** |
| Arbetslösa | 0.2109136 | 0.5011009 | 0.421 | 0.6738 |
| ‘Kommungrupp kod‘2 | 0.4814753 | 0.4093707 | 1.176 | 0.2395 |
| ‘Kommungrupp kod‘3 | 0.7368409 | 0.4110389 | 1.793 | 0.0730 . |
| ‘Kommungrupp kod‘4 | 0.4297422 | 0.4125753 | 1.042 | 0.2976 |
| ‘Kommungrupp kod‘5 | 0.4523037 | 0.4128870 | 1.095 | 0.2733 |
| ‘Kommungrupp kod‘6 | 0.5315953 | 0.4119944 | 1.290 | 0.1969 |
| ‘Kommungrupp kod‘7 | 0.4583103 | 0.4123013 | 1.112 | 0.2663 |
| ‘Kommungrupp kod‘8 | 0.5200929 | 0.4125765 | 1.261 | 0.2075 |
| ‘Kommungrupp kod‘9 | 0.6021329 | 0.4124182 | 1.460 | 0.1443 |
| Utbildad | 0.4874018 | 0.3574511 | 1.364 | 0.1727 |
| ‘MedelInkomst(Tkr)‘ | 0.0005954 | 0.0009268 | 0.642 | 0.5206 |
| Inpendling | -0.1994394 | 0.1723393 | -1.157 | 0.2472 |
| Utpendling | 0.3566750 | 0.1521336 | 2.344 | 0.0191 * |
| (Phi) | 301.70 | 25.55 | 11.81 | 2e-16 |

Pseudo $R^2 = 0.9532$, AIC $= -1365.375$

### 4.1.2 Reducerad modell $M_{RBeta}$

I Tabell 5 visar det starkt pseudo $R^2$ för $M_{Beta}$, däremot är många parametrar inte signifikanta i hypotestestet. Då vi söker den enklaste modellen utförs stegvis selektion för att förenkla modellen.

Tabell 6: Exkluderade variabler från BetaModell

| |
|---|
| Kommungrupp kod |
| Arbetslösa |
| Medelinkomst |
| Utpendling |
| Inpendling |

Efter Stegvis selektion enligt Tabell 6 definieras nya reducerade modellen $M_{RBeta}$.

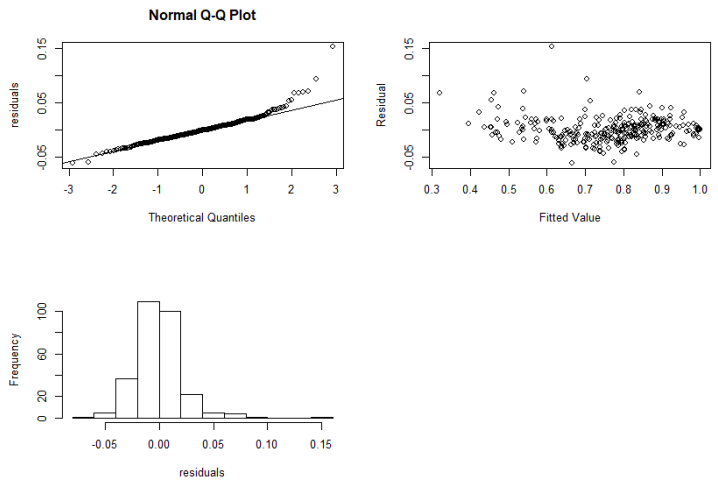
$$g(\mu_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \tag{16}$$

Tabell 7: $M_{RBeta}$

| $M_{RBeta}$ | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.43861 | 0.07768 | -31.394 | 2e-16 *** |
| TätortAndel | 6.45915 | 0.31978 | 20.199 | 2e-16 *** |
| Urbaniserat2010 | 4.14278 | 0.08205 | 50.492 | 2e-16 *** |
| Utbildad | 1.38842 | 0.24183 | 5.741 | 9.39e-09 *** |
| (Phi) | 268.62 | 22.75 | 11.81 | 2e-16 *** |

Pseudo $R^2 = 0.9494$, AIC $= -1382.047$

Tabell 7 visar ingen förbättring eller märkbart förlust med dess $R^2$ mått mot $M_{Beta}$, intressant att notera att *Utbildad* var den enda regressor som skiljde sig i värde från dess lika i $M_{Beta}$ när de övriga tre faktorer är relativt oförändrat.




Figur 14: Residual plot för *Beta Modell*

Med ett mindre AIC värde för $M_{RBeta}$ och med praktiskt samma förklaringsgrad rekommenderas modellen över $M_{Beta}$.

Med en kontroll koll i Figur 14 ser man att det finns en extremt avvikande punkt. Observationen kan exkluderas som outlier, vilket ger en mycket starkare residual plot men ingen markant skillnad i dess koefficienter. Vilket är varför modellen är funnen i Appendix A.2 då punkten hade ingen tydligt påverkan på $M_{RBeta}$.

### 4.1.3 Samspels modell $M_{Sam}$

I en modell kan variablerna ha annorlunda effekt när de kombineras med en annan variabel, detta kallas *samspel* mellan variabler när en av de variabler står för sig själv kallas det huvudeffekt. När minsta möjliga modell är huvud målen vill vi även hitta alla möjliga faktorer som kan ha en roll i tätortsgraden, vi väljer då att söka efter alla möjliga samspel mellan två variabler. Det kan finnas samspel med flera än två variabler men komplexitetet ökas expontientellt med varje grad.

För att hitta andragrads samspel mellan variablerna, använder vi av en modell som har alla möjliga samspel för två variabler och notera vilka kombinationer har signifikant p-värde. För betaregression fann då följande samspel med anseende på p-värde.

I modellen definieras samspelet som:

$$\mu = \alpha + ... + \beta_k x_i x_j ... \tag{17}$$

Tabell 8: Samspel

| | |
|---|---|
| TätortAndel | Urbaniserat2010 |
| TätortAndel | Utbildad |
| Urbaniserat | Utpendling |
| Utbildad | Inpendling |

Där $i, j$ är index för de förklarande variabler och $\beta_k$ är skattning för samspelet, modellen $M_{Sam}$ blev efter stegvis selektion:

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_3 + \beta_3 x_2 x_3 \tag{18}$$

Där $x_1, x_2, x_3$ är värderna för *TätortAndel*, *Urbaniserat2010* och *Utbildad* respektive.

Tabell 9: $M_{Sam}$

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.52452 | 0.07579 | -33.310 | 2e-16 *** |
| TätortAndel | 1.19378 | 2.88804 | 0.413 | 0.679 |
| Urbaniserat2010 | 3.97833 | 0.08868 | 44.860 | 2e-16 *** |
| Utbildad | 2.08091 | 0.27286 | 7.626 | 2.41e-14 *** |
| TätortUrban2010 | 13.57122 | 3.45075 | 3.933 | 8.40e-05 *** |
| TätortUtbildad | -18.30529 | 4.61539 | -3.966 | 7.30e-05 *** |

Pseudo $R^2 = 0.9441$, AIC $= -1396.282$

Det är intressant att notera hur lite det skiljer sig mot $M_{RBeta}$ där samma tre variabler används i slutändan för båda modeller. Dock när $M_{RBeta}$ redan har hög $R^2$ är det nog förväntat att det är minimala förändringar inom förklaringsgraden med $M_{Sam}$.

I jämförelse med $M_{RBeta}$ har vi mindre AIC värde, men på grund av huvudeffektet *TätortAndel* höga p-värde och att $M_{Sam}$ har flera $\beta$ parametrar medför att modellen är mer komplicerad än $M_{RBeta}$ med några svaga förklarande variabler.

QQ plot för $M_{Sam}$ har i princip samma egenskaper som QQplot för $M_{RBeta}$, grafen är därför funnen i Appendix A.2.

## 4.2 LogModell

I avsnitt 3 påträffades en linjär korrelation för logaritmiskt *TätortAndel* och *Urbaniserat2010*. Eftersom kolineäritet medför onödigt komplexitet anses det vara klok att utforska ämnet, en ny modell skapas då en förklarande variabel *TätortAndel* är ersatt av logaritmerat *TätortAndel*.

### 4.2.1 Grund modell $M_{Log}$

Uppsatt exakt som $M_{Beta}$ från föregående sektion så får modellen $M_{Log}$ följande parametrar med *log(TätortAndel)*:

Tabell 10: $M_{Log}$

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -1.592771 | 0.641111 | -2.484 | 0.012977 * |
| Log(TätortAndel) | 0.092335 | 0.022674 | 4.072 | 4.65e-05 *** |
| Urbaniserat2010 | 4.505613 | 0.257508 | 17.497 | ¡ 2e-16 *** |
| Arbetslösa | 2.341623 | 0.878732 | 2.665 | 0.007704 ** |
| ‘Kommungrupp kod‘2 | -1.681660 | 0.516137 | -3.258 | 0.001121 ** |
| ‘Kommungrupp kod‘3 | -1.808469 | 0.511859 | -3.533 | 0.000411 *** |
| ‘Kommungrupp kod‘4 | -2.013603 | 0.515409 | -3.907 | 9.35e-05 *** |
| ‘Kommungrupp kod‘5 | -1.985785 | 0.515239 | -3.854 | 0.000116 *** |
| ‘Kommungrupp kod‘6 | -1.950111 | 0.512365 | -3.806 | 0.000141 *** |
| ‘Kommungrupp kod‘7 | -2.003035 | 0.514344 | -3.894 | 9.85e-05 *** |
| ‘Kommungrupp kod‘8 | -1.800355 | 0.516656 | -3.485 | 0.000493 *** |
| ‘Kommungrupp kod‘9 | -1.724683 | 0.517133 | -3.335 | 0.000853 *** |
| Utbildad | 1.055215 | 0.631340 | 1.671 | 0.094645 . |
| ‘MedelInkomst(Tkr)‘ | 0.004756 | 0.001560 | 3.049 | 0.002292 ** |
| Inpendling | 0.985134 | 0.304090 | 3.240 | 0.001197 ** |
| Utpendling | 0.400725 | 0.278802 | 1.437 | 0.150629 |
| (Phi) | 94.830 | 8.012 | 11.84 | 2e-16 |

Pseudo $R^2 = 0.7982$, AIC $= -1065.567$

I jämförelse med $M_{Beta}$ så har $M_{Log}$ mindre förklaringsgrad och dess AIC i jämförelse är större än Tabell 5. Likadant som $M_{Beta}$ är det många icke signifikanta förklarande variabler, intressant nog när stegvis selektion upprepas exkluderas andra variabler jämfört med $M_{Beta}$.

| Borttagna variabler |
|---|
| Arbetslösa |
| Utbildad |
| Utpendling |

### 4.2.2 Reducerad modell $M_{RLog}$

Med exkluderade variabler enligt Tabell 4.2.1, kan då den nya modellen $M_{RLog}$ definieras:

$$g(\mu_i) = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \beta_5 x_{5j} \tag{19}$$

Tabell 11: $M_{RLog}$

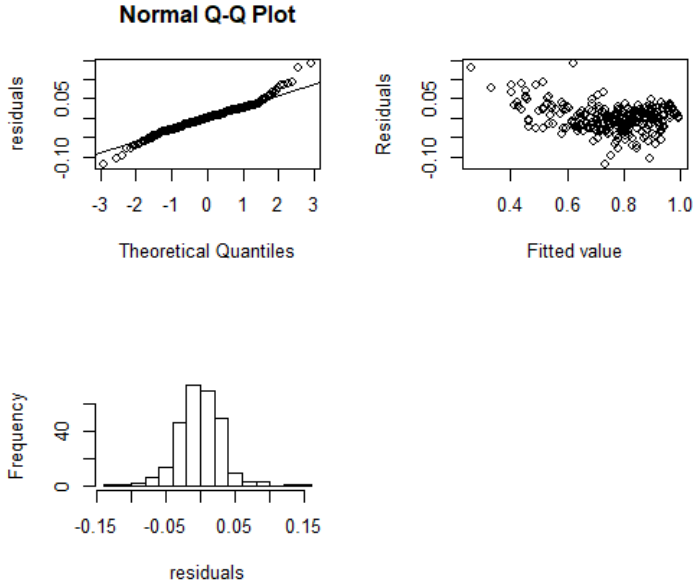
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -0.798211 | 0.593152 | -1.346 | 0.178396 |
| Log(TätortAndel) | 0.102704 | 0.022299 | 4.606 | 4.11e-06 *** |
| Urbaniserat2010 | 4.398538 | 0.179776 | 24.467 | ¡ 2e-16 *** |
| ‘Kommungrupp kod‘2 | -1.725157 | 0.509235 | -3.388 | 0.000705 *** |
| ‘Kommungrupp kod‘3 | -1.829017 | 0.512383 | -3.570 | 0.000357 *** |
| ‘Kommungrupp kod‘4 | -2.080709 | 0.508732 | -4.090 | 4.31e-05 *** |
| ‘Kommungrupp kod‘5 | -2.100274 | 0.510270 | -4.116 | 3.85e-05 *** |
| ‘Kommungrupp kod‘6 | -2.053208 | 0.510314 | -4.023 | 5.74e-05 *** |
| ‘Kommungrupp kod‘7 | -2.101963 | 0.508652 | -4.132 | 3.59e-05 *** |
| ‘Kommungrupp kod‘8 | -1.904034 | 0.513097 | -3.711 | 0.000207 *** |
| ‘Kommungrupp kod‘9 | -1.861370 | 0.514126 | -3.620 | 0.000294 *** |
| ‘MedelInkomst(Tkr)‘ | 0.005024 | 0.001120 | 4.486 | 7.27e-06 *** |
| Inpendling | 0.987200 | 0.306521 | 3.221 | 0.001279 ** |
| (Phi) | 92.937 | 7.852 | 11.84 | 2e-16 *** |

Pseudo $R^2 = 0.7984$, AIC $= -1084.137$

Då anpassningsmåttet för $M_{RLog}$ är lägre än $M_{RBeta}$ i både $R^2$ och AIC, är det mer intressant att se hur variablerna spelas runt *TätortAndel*. Vi noterar vilka variabler som är signifikanta i jämförelse med $M_{RBeta}$.

En sak att notera om $M_{Rlog}$, dess kolineäritet med responsen *Urbaniserat2015* implicerar också korrelation med förklarande variabel *Urbaniserat2010* då både variabler är relativt lika. Om *TätortAndel* exkluderas med avseende på kolineäritet till en förklarande variabel, behålls nästan alla p-värde för faktorerna med ingen markant ändring för varken AIC och förklaringsgrad. Det i kombination med observationen i $M_{RLog}$ ger en intressant diskussions punkt senare i avsnitt 5. Modellen är hittas i Appendix A.2.

Jämför man Figur 15 med Figur 14, så har graferna från Figur 15 inte samma grad av normalfördelning, det har en märkbart skevning åt både extrema värde av residualen med större utbred av dess outlier i jämförelse med $M_{RBeta}$.




Figur 15: Residual plot för *Red LogModell*

## 4.3 Förändrings modell

En gemensamt egenskap hos modellerna visar att *Urbaniserat2010* har stor signifikans för modellen, vi misstänker att *Urbanisering2010* spelar en stor roll för *Urbanisering2015* då bör rimligtvis *Urbaniserat2010* vara relativt lik *Urbaniserat2015*. Frågan är nu om man kan undvika dessa två parametrar i modellen och ändå får liknade resultat till föregående modeller.

Detta undersöks genom att hänvisa till den procentuella förändring mellan År 2015 och 2010 som responsvariabel, vilket ger en model som modellerar förändringen av urbaniseringen mellan perioderna. Det finns två sätt att tolka förändringen, första exempel där responsvariablerna är den totala förändringen som är differensen av tätortsgraden mellan 2015 och 2010 och relativa förändring vilket är kvoten mellan 2015 och 2010 alltså procentuell tillväxt.

Eftersom responsvariabeln nu inte är en värde inom [0,1], används multipel linjär regression. Detta tjänar också som en jämförande blick med en icke beta-regression.

### 4.3.1 Deltamodell $M_D$

Modellen för totala förändringen, $M_D$, definieras:

$$\mu_i = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \beta_5 x_{5j} + \beta_6 x_{6j} + \beta_7 x_{7j} \tag{20}$$

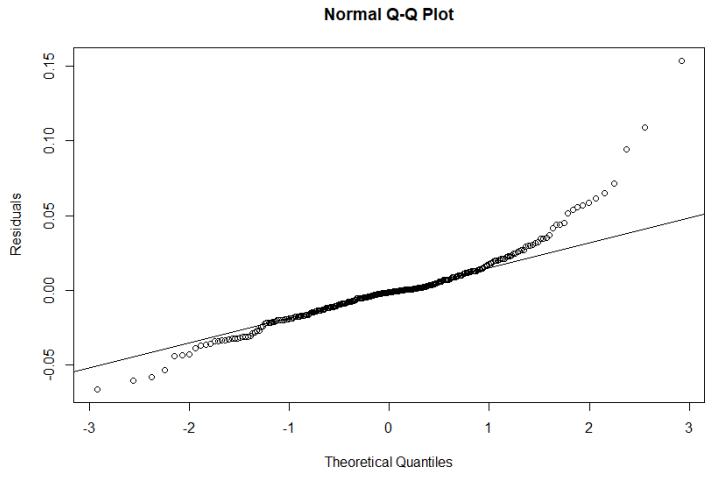
Där responsvariabeln, $Y_{delta} = x_{Urbaniserat2015} - x_{Urbaniserat2010}$

Tabell 12: DeltaModell

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -0.0234 | 0.0324 | -0.72 | 0.4715 |
| TätortAndel | -0.0793 | 0.0216 | -3.67 | 0.0003 |
| Arbetslösa | 0.0045 | 0.0880 | 0.05 | 0.9589 |
| ‘Kommungrupp kod‘2 | -0.0302 | 0.0188 | -1.61 | 0.1087 |
| ‘Kommungrupp kod‘3 | -0.0255 | 0.0189 | -1.35 | 0.1784 |
| ‘Kommungrupp kod‘4 | -0.0338 | 0.0195 | -1.74 | 0.0830 |
| ‘Kommungrupp kod‘5 | -0.0303 | 0.0196 | -1.55 | 0.1229 |
| ‘Kommungrupp kod‘6 | -0.0171 | 0.0191 | -0.90 | 0.3700 |
| ‘Kommungrupp kod‘7 | -0.0242 | 0.0195 | -1.24 | 0.2161 |
| ‘Kommungrupp kod‘8 | -0.0204 | 0.0194 | -1.05 | 0.2939 |
| ‘Kommungrupp kod‘9 | 0.0012 | 0.0199 | 0.06 | 0.9520 |
| Utbildad | 0.1087 | 0.0549 | 1.98 | 0.0487 |
| ‘MedelInkomst(Tkr)‘ | 0.0000 | 0.0001 | 0.35 | 0.7288 |
| Inpendling | 0.0038 | 0.0243 | 0.15 | 0.8771 |
| Utpendling | 0.1137 | 0.0208 | 5.47 | 0.0000 |

$R^2$ = 0.2489, AIC = -1252.056

Tabell 12 visar $M_D$ har lägt $R^2$ och att det finns en skevning i Figur 16 säger att modellen har lite anpassning till data, därför väljer vi inte gå vidare med modellen.




Figur 16: QQplot för DeltaModell

### 4.3.2 Relativmodell $M_R$

Relativa förändring eller $M_R$ är lik $M_D$, det som skiljer är responsvariabeln. Förutom det använder $M_R$ linjär regression som $M_D$:

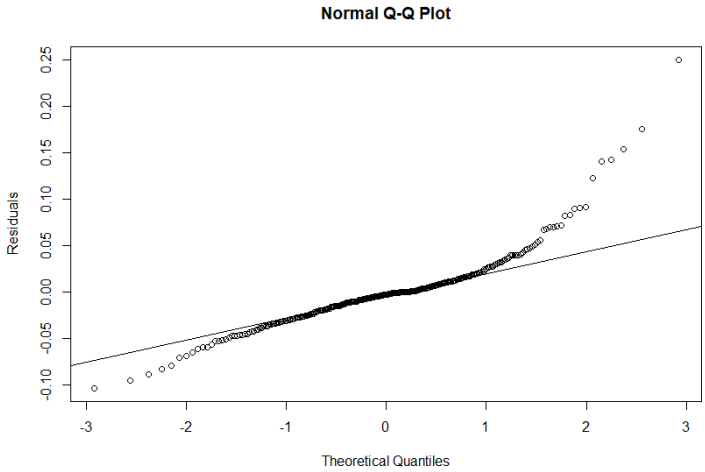
$$Y_{Relativt} = x_{Urbaniserat2010}/x_{Urbaniseratg2015} \tag{21}$$

Tabell 13: $M_R$

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.9661 | 0.0528 | 18.31 | 0.0000 |
| TätortAndel | -0.1011 | 0.0352 | -2.87 | 0.0044 |
| Arbetslösa | 0.0250 | 0.1432 | 0.17 | 0.8614 |
| ‘Kommungrupp kod‘2 | -0.0652 | 0.0306 | -2.13 | 0.0338 |
| ‘Kommungrupp kod‘3 | -0.0327 | 0.0308 | -1.06 | 0.2895 |
| ‘Kommungrupp kod‘4 | -0.0625 | 0.0317 | -1.97 | 0.0494 |
| ‘Kommungrupp kod‘5 | -0.0439 | 0.0319 | -1.38 | 0.1703 |
| ‘Kommungrupp kod‘6 | -0.0201 | 0.0311 | -0.65 | 0.5190 |
| ‘Kommungrupp kod‘7 | -0.0410 | 0.0318 | -1.29 | 0.1980 |
| ‘Kommungrupp kod‘8 | -0.0223 | 0.0316 | -0.71 | 0.4802 |
| ‘Kommungrupp kod‘9 | 0.0108 | 0.0323 | 0.33 | 0.7394 |
| Utbildad | 0.1386 | 0.0894 | 1.55 | 0.1221 |
| ‘MedelInkomst(Tkr)‘ | -0.0000 | 0.0002 | -0.01 | 0.9889 |
| Inpendling | 0.0140 | 0.0395 | 0.35 | 0.7230 |
| Utpendling | 0.2609 | 0.0339 | 7.70 | 0.0000 |

$R^2 = 0.2984$, AIC $= -979.2086$

Tabell 13 visar ingen imponerande $R^2$ även i jämförelse till $M_D$.




Figur 17: QQPlot för relativt förändring

Figur 17 visar samma karaktär som Figur 16, vilket indikerar att ingen av modellerna följer inte normalfördelning och därmed uppfyller inte kraven för linjär regression.

Från dessa resultat dra vi slutsatsen att den nuvarande urbaniseringsgraden har en stor effekt på modellen och att modellera utan den variabeln behövs något annat för att främja ett vettigt modell, att vi också använder linjär regression kan vara delaktigt med modellens dåliga anpassning till data.

I alla fall ser det inte övertygande ut för någon av modellerna därmed så förkastas variabler selektion för modellerna då ingen tillfredställande $R^2$ skulle till komma från så låga anpassningsmått.

# 5 Slutsats

I slutet anses att $M_{RBeta}$ bäst uppfylla frågeställningen angiven i avsnitt 1.1:

$$g(\mu) = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \epsilon_{ij} \tag{22}$$

Tabell 14: $M_{RBeta}$

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.43861 | 0.07768 | -31.394 | 2e-16 *** |
| TätortAndel | 6.45915 | 0.31978 | 20.199 | 2e-16 *** |
| Urbaniserat2010 | 4.14278 | 0.08205 | 50.492 | 2e-16 *** |
| Utbildad | 1.38842 | 0.24183 | 5.741 | 9.39e-09 *** |
| (Phi) | 268.62 | 22.75 | 11.81 | 2e-16 *** |

Det är lika vettigt att använda $M_{Sam}$ som slut modell, däremot med avseende till frågeställningen i avsnitt 1.1 var målen att hitta den minst komplicerade modellen. I jämförelse mot $M_{RBeta}$ har $M_{Sam}$ fler variabler utan något markant skillnad i $R^2$ eller AIC. Dessutom är $M_{RBeta}$ mycket enklare att tolka och delar samma förklarande variabler med $M_{Sam}$.

Tabell 15: Test av anpassning

| Kommun | Obsverad värde | Skattad värde | residual fel $(y - \hat{y})$ |
|---|---|---|---|
| 1861 Hallsberg | 0.7885 | 0.7767 | 0.011 |
| 1781 Kristinehamn | 0.8367 | 0.8159 | 0.020 |
| 1762 Munkfors | 0.7712 | 0.8315 | 0.060 |
| 1881 Kumla | 0.8744 | 0.8764 | 0.001 |
| 2404 Vindeln | 0.5924 | 0.6023 | 0.009 |
| 2313 Strömsund | 0.5804 | 0.5714 | 0.009 |
| 0685 Vetlanda | 0.7671 | 0.7692 | 0.002 |
| 2580 Luleå | 0.9035 | 0.8938 | 0.009 |
| 0780 Växjö | 0.8680 | 0.8830 | 0.014 |
| 1293 Hässleholm | 0.8074 | 0.8203 | 0.012 |

## 5.1 Tolkning av modellen

Från $M_{RBeta}$ och de övriga modeller kan man notera ett en gemensam egenskap hos de modeller med höga anpassingsmått, dessa modeller delar två gemensamma förklarande variabler med varandra, *Urbaniserat* och *TätortAndel*. Med denna observation kan vi tolka från modellerna att dessa två faktorer är mycket viktiga för urbaniseringsgraden.

Resultatet från $M_D$ och $M_R$ stödjer detta påstående då dessa modeller var ett försök att utföra regression utan *Urbaniserat2010* vilket påvisa att den nuvarande urbaniseringsgrad spelar en stor roll i modellering då varken $M_D$ och $M_R$ hade ingen tillfredställande anpassningsmått.

*TätortAndel* har en stor effekt på modellerna, i $M_{RLog}$ noterade vi att med logaritmiskt *TätortAndel* och dess kolineäritet med responsvariablern gav $M_{RLog}$ helt andra egenskaper med dess förklarande variabler i jämförelse mot $M_{RBeta}$.

Variabeln *Utbildad* i $M_{RBeta}$ kunde på samma sätt tas bort utan förlust av $R^2$ eller AIC säger hur mycket effekt *Urbaniserat2010* och *TätortAndel* har i modellerna. Bara när modellen består enbart *TätortAndel* eller *Urbaniserat2010* finner man en signifikant förlust av förklaringsgrad.

Från det kan man dra slutsatsen att man kan rimligen uppskatta kommunens framtida tätortsgrad enbart från dess Tätortsareal och befolkning i tätorten utan större hänsyn till andra faktorer. Ingen andra variabler spelar lika stor förklarande roll som dessa två faktorer, man kan säga att urbaniseringen livnär sig själv. Högt tätbebyggt land eller befolkning leder till växande tätorter, vilket är förnuftigt men personligen antog man att nån annan faktor hade spelat roll. Även observeras att tätortsgraden växer exponentiellt med *TätortAndel* som Figur 12 visar.

En enkel slutsats, men som Tabell 15 visar är det skattade värderna för test kommunerna lovande nära den verkliga värde, så målet satt för uppsatsen är nått.

För själva modellen $M_{RBeta}$, i Tabell 14, ser man att *TätortAndel* i kommunen är parametern med högst värde. Det implicerar att varje värde av tätbebyggt mark leder till mer urbanisering än av befolkningen redan i tätorten, med utbildningsbivå som svagaste faktor i urbaniseringen.

Eftersom *Red BetaModell* visar starka anpassningsmått utförde prediktion av framtida tätortsgrad för alla kommuner, den tabellen finns att hitta i Appendix A.1.

# 6 Diskussion

En grundläggand antagande är att varje observation av data är oberoende av varandra. I detta uppsats togs den antagelse för givet, men det kan vara en bidragande faktor.

Som $M_{RLog}$ visar har de utelämnade variablerna *Kommungrupp kod* och *Inpendling* en effekt på tätortsgraden, men i jämförelse mot $M_{RBeta}$ har $M_{RLog}$ lägre anpassningsmått.
Med hänsyn till Occams Rakniv påstår vi att det är inte nödvändigt att fortsätta med tanken då vi redan har en simpel modell som förklarar urbanisering godtyckligt och flera variabler kommer inte ge något mer till modellen redan höga anpassningsmått. Dessutom är det enklare att samla data om relativt enkla mätbara parametrar (befolkning, area) än mer komplicerat faktorer (Utbildning nivå, pendling, inkomst och resterande faktorer).

Med andra variabler, kan det finnas faktorer som vi inte tagit hänsyn till i studien, om det verkligen kan påverka modellen är tveksamt på grund av dess redan starka förklaringsgrad.

Eftersom vi också enbart baserat vår data på två tidpunkter kan det vara möjligt att modellen bara är giltig mellan dessa två perioder, trots allt kan samhälle, attityd eller policy vara faktorer som inte är helt lätta kvantifiera en modell. Frågan är också om hur mycket det kan spelar roll i vår valda modell.

I början hänvisade det till en studie [7] om tillväxt, de föreslog att transportation och koppling till andra centrum, utvecklad näringsliv och utbildning som faktorer till god tillväxt. Det är överraskande att ingen av de nämnda faktorer var signifikanta i vår modell. Endast en faktor *Utbildad* blev signifikant och det i sig är nog inte nödvändigt för modellen i jämförelse med *Urbaniserat2010* eller *TätortAndel*

En liknade studie[8], som studerade tillväxt i kommunnivå noterade att motivering till en typ av tillväxt som en faktor, dessa involverar specifika initiativ vilket är inte konkreta värde att anpassa till regressionsmodell. De noterar också att definitionen av tillväxt är upp till kommunens egna intresse, och detta intresse nödvändigtvis inte relaterat till urbaniseringen av kommunen.

## 6.1 Fortsätt arbete

Försöken med att modellera förändring syftades i grunden att visa att nuvarande urbaniseringsgrad har en stor effekt på den framtida tätortsgrad, men det har en intressant fortsättning. En djupare titt på detta modell kan ge en annan insikt på hur kommunernas urbaniseringsgrad förändras mellan perioderna istället vad den blir i den andra perioden.

# 7 Appendix

## 7.1 A.1

Predikterad 2020 Urbaniseringgrad baserad på år 2015 parametrar.

| region | Urbanisering | region | Urbanisering |
|---|---|---|---|
| 0114 Upplands Väsby | 0.97 | 0115 Vallentuna | 0.88 |
| 0117 Österåker | 0.95 | 0120 Värmdö | 0.96 |
| 0123 Järfälla | 0.99 | 0125 Ekerö | 0.89 |
| 0126 Huddinge | 0.99 | 0127 Botkyrka | 0.96 |
| 0128 Salem | 0.97 | 0136 Haninge | 0.95 |
| 0138 Tyresö | 1.00 | 0139 Upplands-Bro | 0.90 |
| 0140 Nykvarn | 0.83 | 0160 Täby | 1.00 |
| 0162 Danderyd | 1.00 | 0163 Sollentuna | 1.00 |
| 0180 Stockholm | 1.00 | 0181 Södertälje | 0.92 |
| 0182 Nacka | 1.00 | 0183 Sundbyberg | 1.00 |
| 0184 Solna | 1.00 | 0186 Lidingö | 1.00 |
| 0187 Vaxholm | 0.97 | 0188 Norrtälje | 0.71 |
| 0191 Sigtuna | 0.89 | 0192 Nynäshamn | 0.86 |
| 0305 Håbo | 0.93 | 0319 Älvkarleby | 0.89 |
| 0330 Knivsta | 0.79 | 0331 Heby | 0.65 |
| 0360 Tierp | 0.68 | 0380 Uppsala | 0.90 |
| 0381 Enköping | 0.74 | 0382 Östhammar | 0.72 |
| 0428 Vingåker | 0.74 | 0461 Gnesta | 0.72 |
| 0480 Nyköping | 0.85 | 0481 Oxelösund | 0.99 |
| 0482 Flen | 0.77 | 0483 Katrineholm | 0.83 |
| 0484 Eskilstuna | 0.90 | 0486 Strängnäs | 0.83 |
| 0488 Trosa | 0.88 | 0509 Ödeshög | 0.58 |
| 0512 Ydre | 0.49 | 0513 Kinda | 0.70 |
| 0560 Boxholm | 0.72 | 0561 Åtvidaberg | 0.79 |
| 0562 Finspång | 0.82 | 0563 Valdemarsvik | 0.65 |
| 0580 Linköping | 0.91 | 0581 Norrköping | 0.90 |
| 0582 Söderköping | 0.71 | 0583 Motala | 0.85 |
| 0584 Vadstena | 0.80 | 0586 Mjölby | 0.84 |
| 0604 Aneby | 0.66 | 0617 Gnosjö | 0.77 |
| 0642 Mullsjö | 0.87 | 0643 Habo | 0.81 |
| 0662 Gislaved | 0.83 | 0665 Vaggeryd | 0.78 |
| 0680 Jönköping | 0.90 | 0682 Nässjö | 0.84 |
| 0683 Värnamo | 0.82 | 0684 Sävsjö | 0.77 |
| 0685 Vetlanda | 0.79 | 0686 Eksjö | 0.81 |
| 0687 Tranås | 0.86 | 0760 Uppvidinge | 0.75 |
| 0761 Lessebo | 0.86 | 0763 Tingsryd | 0.67 |
| 0764 Alvesta | 0.73 | 0765 Älmhult | 0.74 |
| 0767 Markaryd | 0.79 | 0780 Växjö | 0.88 |
| 0781 Ljungby | 0.75 | 0821 Högsby | 0.69 |
| 0834 Torsås | 0.62 | 0840 Mörbylånga | 0.82 |
| 0860 Hultsfred | 0.81 | 0861 Mönsterås | 0.78 |
| 0862 Emmaboda | 0.74 | 0880 Kalmar | 0.90 |
| 0881 Nybro | 0.80 | 0882 Oskarshamn | 0.84 |

| | | | |
|---|---|---|---|
| 0883 Västervik | 0.82 | 0884 Vimmerby | 0.76 |
| 0885 Borgholm | 0.58 | 0980 Gotland | 0.69 |
| 1060 Olofström | 0.83 | 1080 Karlskrona | 0.88 |
| Ronneby | 0.84 | 1082 Karlshamn | 0.88 |
| 1083 Sölvesborg | 0.91 | 1214 Svalöv | 0.75 |
| 1230 Staffanstorp | 0.93 | 1231 Burlöv | 0.99 |
| 1233 Vellinge | 0.96 | 1256 Östra Göinge | 0.81 |
| 1257 Örkelljunga | 0.79 | 1260 Bjuv | 0.93 |
| 1261 Kävlinge | 0.95 | 1262 Lomma | 0.97 |
| 1263 Svedala | 0.88 | 1264 Skurup | 0.83 |
| 1265 Sjöbo | 0.73 | 1266 Hörby | 0.67 |
| 1267 Höör | 0.86 | 1270 Tomelilla | 0.75 |
| 1272 Bromölla | 0.89 | 1273 Osby | 0.80 |
| 1275 Perstorp | 0.83 | 1276 Klippan | 0.80 |
| 1277 Åstorp | 0.93 | 1278 Båstad | 0.88 |
| 1280 Malmö | 1.00 | 1281 Lund | 0.95 |
| 1282 Landskrona | 0.95 | 1283 Helsingborg | 0.96 |
| 1284 Höganäs | 0.95 | 1285 Eslöv | 0.86 |
| 1286 Ystad | 0.88 | 1287 Trelleborg | 0.89 |
| 1290 Kristianstad | 0.88 | 1291 Simrishamn | 0.81 |
| 1292 Ängelholm | 0.90 | 1293 Hässleholm | 0.84 |
| 1315 Hylte | 0.70 | 1380 Halmstad | 0.92 |
| 1381 Laholm | 0.78 | 1382 Falkenberg | 0.83 |
| 1383 Varberg | 0.86 | 1384 Kungsbacka | 0.93 |
| 1401 Härryda | 0.94 | 1402 Partille | 0.99 |
| 1407 Öckerö | 0.99 | 1415 Stenungsund | 0.91 |
| 1419 Tjörn | 0.90 | 1421 Orust | 0.56 |
| 1427 Sotenäs | 0.90 | 1430 Munkedal | 0.66 |
| 1435 Tanum | 0.49 | 1438 Dals-Ed | 0.64 |
| 1439 Färgelanda | 0.61 | 1440 Ale | 0.87 |
| 1441 Lerum | 0.95 | 1442 Vårgårda | 0.59 |
| 1443 Bollebygd | 0.69 | 1444 Grästorp | 0.57 |
| 1445 Essunga | 0.50 | 1446 Karlsborg | 0.81 |
| 1447 Gullspång | 0.70 | 1452 Tranemo | 0.74 |
| 1460 Bengtsfors | 0.72 | 1461 Mellerud | 0.62 |
| 1462 Lilla Edet | 0.72 | 1463 Mark | 0.74 |
| 1465 Svenljunga | 0.64 | 1466 Herrljunga | 0.60 |
| 1470 Vara | 0.62 | 1471 Götene | 0.70 |
| 1472 Tibro | 0.85 | 1473 Töreboda | 0.65 |
| 1480 Göteborg | 0.99 | 1481 Mölndal | 0.98 |
| 1482 Kungälv | 0.88 | 1484 Lysekil | 0.83 |
| 1485 Uddevalla | 0.86 | 1486 Strömstad | 0.71 |
| 1487 Vänersborg | 0.84 | 1488 Trollhättan | 0.92 |
| 1489 Alingsås | 0.88 | 1490 Borås | 0.92 |
| 1491 Ulricehamn | 0.73 | 1492 Åmål | 0.81 |
| 1493 Mariestad | 0.81 | 1494 Lidköping | 0.82 |
| 1495 Skara | 0.80 | 1496 Skövde | 0.89 |
| 1497 Hjo | 0.77 | 1498 Tidaholm | 0.75 |
| 1499 Falköping | 0.78 | 1715 Kil | 0.80 |

| | | | | | |
|---|---|---|---|---|---|
| 1730 | Eda | 0.57 | 1737 | Torsby | 0.49 |
| 1760 | Storfors | 0.64 | 1761 | Hammarö | 0.97 |
| 1762 | Munkfors | 0.81 | 1763 | Forshaga | 0.84 |
| 1764 | Grums | 0.75 | 1765 | Årjäng | 0.48 |
| 1766 | Sunne | 0.54 | 1780 | Karlstad | 0.91 |
| 1781 | Kristinehamn | 0.84 | 1782 | Filipstad | 0.76 |
| 1783 | Hagfors | 0.68 | 1784 | Arvika | 0.71 |
| 1785 | Säffle | 0.68 | 1814 | Lekeberg | 0.57 |
| 1860 | Laxå | 0.74 | 1861 | Hallsberg | 0.81 |
| 1862 | Degerfors | 0.81 | 1863 | Hällefors | 0.76 |
| 1864 | Ljusnarsberg | 0.73 | 1880 | Örebro | 0.90 |
| 1881 | Kumla | 0.90 | 1882 | Askersund | 0.66 |
| 1883 | Karlskoga | 0.90 | 1884 | Nora | 0.79 |
| 1885 | Lindesberg | 0.77 | 1904 | Skinnskatteberg | 0.63 |
| 1907 | Surahammar | 0.86 | 1960 | Kungsör | 0.82 |
| 1961 | Hallstahammar | 0.89 | 1962 | Norberg | 0.80 |
| 1980 | Västerås | 0.93 | 1981 | Sala | 0.74 |
| 1982 | Fagersta | 0.88 | 1983 | Köping | 0.84 |
| 1984 | Arboga | 0.84 | 2021 | Vansbro | 0.77 |
| 2023 | Malung-Sälen | 0.73 | 2026 | Gagnef | 0.80 |
| 2029 | Leksand | 0.84 | 2031 | Rättvik | 0.81 |
| 2034 | Orsa | 0.83 | 2039 | Älvdalen | 0.69 |
| 2061 | Smedjebacken | 0.73 | 2062 | Mora | 0.85 |
| 2080 | Falun | 0.87 | 2081 | Borlänge | 0.91 |
| 2082 | Säter | 0.71 | 2083 | Hedemora | 0.74 |
| 2084 | Avesta | 0.84 | 2085 | Ludvika | 0.86 |
| 2101 | Ockelbo | 0.63 | 2104 | Hofors | 0.79 |
| 2121 | Ovanåker | 0.67 | 2132 | Nordanstig | 0.54 |
| 2161 | Ljusdal | 0.64 | 2180 | Gävle | 0.91 |
| 2181 | Sandviken | 0.86 | 2182 | Söderhamn | 0.80 |
| 2183 | Bollnäs | 0.74 | 2184 | Hudiksvall | 0.76 |
| 2260 | Ånge | 0.63 | 2262 | Timrå | 0.84 |
| 2280 | Härnösand | 0.81 | 2281 | Sundsvall | 0.88 |
| 2282 | Kramfors | 0.72 | 2283 | Sollefteå | 0.67 |
| 2284 | Örnsköldsvik | 0.78 | 2303 | Ragunda | 0.47 |
| 2305 | Bräcke | 0.50 | 2309 | Krokom | 0.62 |
| 2313 | Strömsund | 0.60 | 2321 | Åre | 0.67 |
| 2326 | Berg | 0.41 | 2361 | Härjedalen | 0.66 |
| 2380 | Östersund | 0.88 | 2401 | Nordmaling | 0.59 |
| 2403 | Bjurholm | 0.43 | 2404 | Vindeln | 0.62 |
| 2409 | Robertsfors | 0.50 | 2417 | Norsjö | 0.61 |
| 2418 | Malå | 0.68 | 2421 | Storuman | 0.61 |
| 2422 | Sorsele | 0.49 | 2425 | Dorotea | 0.56 |
| 2460 | Vännäs | 0.74 | 2462 | Vilhelmina | 0.55 |
| 2463 | Åsele | 0.62 | 2480 | Umeå | 0.90 |
| 2481 | Lycksele | 0.76 | 2482 | Skellefteå | 0.83 |
| 2505 | Arvidsjaur | 0.77 | 2506 | Arjeplog | 0.65 |
| 2510 | Jokkmokk | 0.75 | 2513 | Överkalix | 0.54 |
| 2514 | Kalix | 0.79 | 2518 | Övertorneå | 0.56 |

| | | | |
|---|---|---|---|
| 2521 Pajala | 0.53 | 2523 Gällivare | 0.82 |
| 2560 Älvsbyn | 0.75 | 2580 Luleå | 0.90 |
| 2581 Piteå | 0.84 | 2582 Boden | 0.84 |
| 2583 Haparanda | 0.78 | 2584 Kiruna | 0.85 |

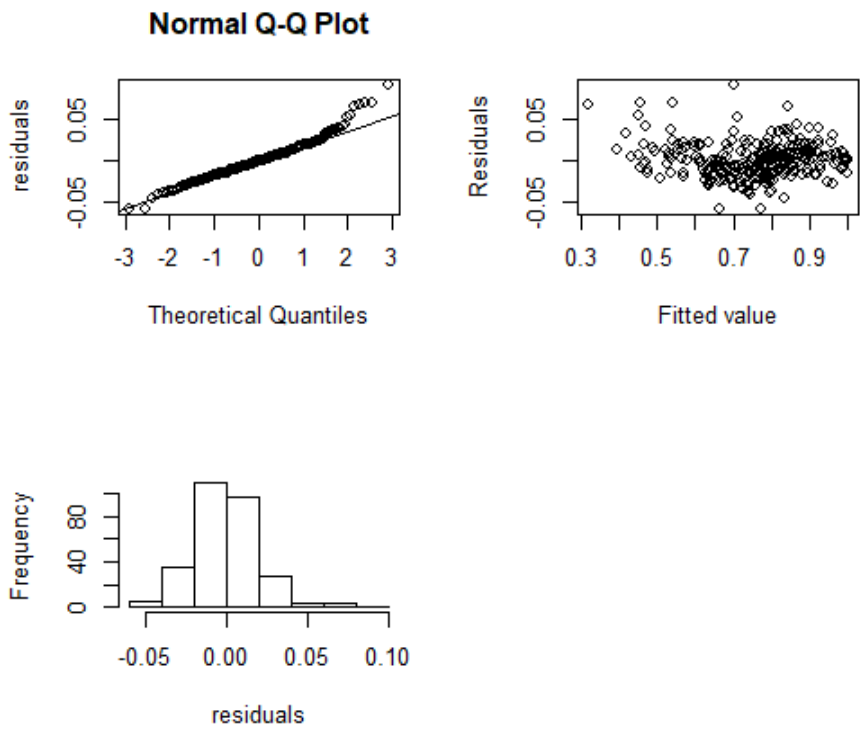
## 7.2 A.2

Samling av ytterligare modeller.

Tabell 17: Reducerad BetaModell utan outlier

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.48013 | 0.07398 | -33.525 | 2e-16 *** |
| TätortAndel | 6.51309 | 0.30889 | 21.085 | 2e-16 *** |
| Urbaniserat | 4.14124 | 0.07760 | 53.366 | 2e-16 *** |
| Utbildad | 1.50784 | 0.22975 | 6.563 | 5.28e-11 *** |
| (Phi) | 301.61 | 25.57 | 11.79 | 2e-16 *** |

Pseudo $R^2 = 0.9503$, AIC $= -1382.569$




Figur 18: QQ och residual plot utan outlier

Tabell 18: BetaModell med bara tätort och urbaniserat2010

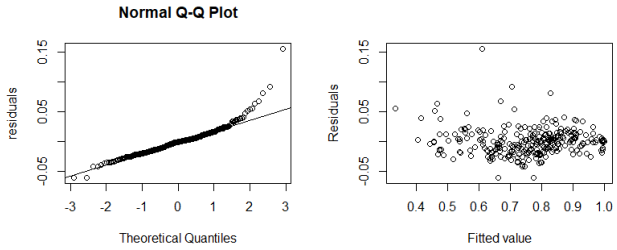
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.10714 | 0.05407 | -38.97 | 2e-16 *** |
| TätortAndel | 7.00884 | 0.32076 | 21.85 | 2e-16 *** |
| Urbaniserat2010 | 4.29127 | 0.08197 | 52.35 | 2e-16 *** |
| (Phi) | 242.61 | 20.56 | 11.8 | 2e-16 *** |

Pseudo $R^2 = 0.9496$, AIC $= -1325.121$

Tabell 19: LogModell utan TätortAndel

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -1.631652 | 0.572950 | -2.848 | 0.004402 ** |
| Urbaniserat | 4.910910 | 0.146090 | 33.616 | ¡ 2e-16 *** |
| ‘Kommungrupp kod‘2 | -1.798275 | 0.510157 | -3.525 | 0.000424 *** |
| ‘Kommungrupp kod‘3 | -1.969001 | 0.512614 | -3.841 | 0.000122 *** |
| ‘Kommungrupp kod‘4 | -2.212983 | 0.508871 | -4.349 | 1.37e-05 *** |
| ‘Kommungrupp kod‘5 | -2.268507 | 0.509859 | -4.449 | 8.62e-06 *** |
| ‘Kommungrupp kod‘6 | -2.203588 | 0.510268 | -4.318 | 1.57e-05 *** |
| ‘Kommungrupp kod‘7 | -2.232365 | 0.508792 | -4.388 | 1.15e-05 *** |
| ‘Kommungrupp kod‘8 | -2.164023 | 0.511207 | -4.233 | 2.30e-05 *** |
| ‘Kommungrupp kod‘9 | -2.091765 | 0.512374 | -4.082 | 4.46e-05 *** |
| ‘MedelInkomst(Tkr)‘ | 0.005614 | 0.001135 | 4.947 | 7.53e-07 *** |
| Inpendling | 1.433140 | 0.299404 | 4.787 | 1.70e-06 *** |
| (Phi) | 85.888 | 7.192 | 11.94 | 2e-16 *** |

Pseudo $R^2 = 0.7901$, AIC $= -1063.305$




Figur 19: QQplot för SamspelsModell

# Referenser


[1] *Sveriges Kommuner och Landsting* Kommunindelning och dess beskrivning (2017)
https://skl.se/tjanster/kommunerregioner/faktakommunerochregioner/kommungruppsindelning.2051.html

[2] *Statistiska CentralByrå* Databyrå för kommun relaterat data och definition av tätorter. (2019)
https://www.scb.se/hitta-statistik/statistik-efter-amne/miljo/markanvandning/tatorter/

[3] *Arbets Förmedlingen* Data för arbetslöshet och definition
https://www.arbetsformedlingen.se/Om-oss/Statistik-och-publikationer/Statistik/Tidigare-statistik.html

[4] *Rolf Sundberg* (2016) Linjära Statistiska Modeller Kompendium Oktober 2016

[5] *Agresti, Alan* (2013) An introduction to Categorical Data Analysis

[6] *Francisco Cribari-Neto, Achim Zeileis* Beta Regression in R.
https://cran.r-project.org/web/packages/betareg/vignettes/betareg.pdf

[7] *Tillväxtanalys* Städer och deras tillväxtförutsättningar (2011)
http://www.tillvaxtanalys.se/download/18.201965214d8715afd14c840/1432715258250/Rapport_2011_08.pdf

[8] *Jonas Fjertorp, Rolf G Larsson, Ola Mattisson* (2012) Kommunaltillväxt, Konsten att hantera lokala förutsättningar
https://www.natkom.se/PDF/Rapporter/007.pdf