Stockholms universitet

Jämförelse mellan regressionsmodeller

Karin Segerros

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2014:14
Matematisk statistik
December 2014

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2014:14**
http://www.math.su.se/matstat

Jämförelse mellan regressionsmodeller

Karin Segerros*

Oktober 2014

Sammanfattning

Inom statistik är regressionsanalys ett viktigt verktyg. Den vanligaste är linjär regression som dock kräver vissa förutsättningar. I detta arbete kommer linjär regression att jämföras med en robust regression då dessa förutsättningar inte är uppfyllda. Resultatet visar att den robusta regressionen är att föredra vid de flesta situationer. Men det visas även att skillnaderna mellan regressionsmodellerna i många av fallen inte är speciellt stora och att den linjära regressionen då skulle kunna vara att föredra pga lättare beräkningar.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: karinsegerros@hotmail.com. Handledare: Gudrun Brattström.

Abstract

An important tool in statistics is regression analysis. Most common is linear regression but it makes some assumtions. This project will compare linear regression with robust regression when the assumtions are not true. The result shows that the robust regression is preferable in most situations. But it will also show that there are small differences between the two regressionsmodels in many of the cases. Thus the linear regression may be better to use because of easier calculations.

Förord

Jag vill tacka min handledare Gudrun Brattstöm för all hjälp. Hennes kunskaper och otroliga tålamod har betytt mycket för detta arbete.

Innehåll

1 Introduktion

Syftet med detta projekt är att jämföra linjär regression med robust regression i situationer där förutsättningarna för linjär regression inte är uppfyllda. Detta kommer göras genom att använda en tidserie X (simulerade temperaturer med hjälp av klimatmodeller), där responsen Y fås genom att addera olika typer brus till X. Bruset kan exempelvis innehålla outliers eller vara autokorrelerat vilket gör att förutsättningarna för linjär regression ej är uppfyllda och vi får en respons där observationerna t ex inte är oberoende eller normalfördelade. Som validering delas datan in i två delar. Först görs en modell på 1:a halvan av datasetet för att skatta det linjära sambandets parametrar. Sedan görs en prediktion med hjälp av modellen för 2:a halvan av datasetet. Dessa predikterade värden jämförs med den faktiska responsen (2:a halvan av datasetet), för att kunna avgöra hur bra modellen är. På detta sätt kan olika modeller jämföras.
För att undvika enstaka tillfälligheter och få en säker analys görs det 4 replikat för varje kombination av regressionsmodell och brus.

2 Beskrivning av data

Datan som används i detta projekt är simulerade data (mha klimatmodeller) för norra halvklotets medeltemperatur. Datan består av 12000 punkter under 1000 år, dvs månadsvisa för åren 1000-1999. Simuleringarna görs i Kelvin men har i detta arbete gjorts om till Celsius.

3 Teori

3.1 Regression

Regressionsanalys är ett viktigt statistiskt verktyg som är vanligt förekommande inom de flesta områden. Det finns olika metoder att använda vid regression. Den vanligaste (och lättaste) är linjär regression.
Den allmänna linjära modellen definieras som

$$Y = \mu + \varepsilon,$$

$$\mu = A\theta, \;\; \varepsilon \sim N(0, \tau^2 I_N)$$

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{N1} & \cdots & a_{Nk} \end{pmatrix}$$

där Y är en stokastisk vektor, μ är väntevärdesvektorn, θ är parametervektorn och ε kallas för försöksfelet. Y, μ och ε har dimension N medan θ har

dimension k. I_N är identitetsmatrisen. (Sundberg. (2012)).
Det finns både fördelar och nackdelar med denna metod.
Fördelen är att parameterskattningarna är lätta att beräkna men nackdelen är att den kräver vissa förutsättningar och att den är känslig för outliers. (Rousseeuw, Leroy. (2005)).
Förutsättningarna för att denna modell ska gälla är att observationerna är oberoende och normalfördelade. (Sundberg. (2012)).
All estimation utgår från antaganden. Metoder som ej är känsliga för avvikelser från modellförutsättningarna kallas för robusta metoder.
(Fox, Weisberg. (2010)).
För en robust estimator gäller att:

- en liten förändring i datan ska inte medföra stora förändringar hos estimatet.
- estimatorn ska vara effektiv under många omständigheter.

(Andersen. (2008)).

Robusta metoder är framförallt väldigt användbara för att upptäcka outliers. För att upptäcka outliers görs en regression på majoriteten av datan för att sedan upptäcka eventuella punkter med höga residualvärden. Den stora skillanden mellan robust regression och outliers detection är att den robusta metoden behåller datapunkten, men låter den inte påverka regressionen medan i outliers detection tar man bort punkterna ur modellen. (Rousseeuw et al).
Den finns olika robusta metoder t ex R-estimation (estimates derived from Rank tests), L-estimation (Linear combinations of order statistics) och M-estimation (Maximum likelihood type). (Huber, Ronchetti, (2009)). I nästföljande avsnitt beskrivs M-estimation mer noggrannt och det är den metod som kommer att användas senare i den robusta regressionsanalysen. Mer bestämt är det Hubers M-estimator som kommer att beskrivas och användas.

3.2 M-estimator

Denna klass av estimatorer kan ses som en generalisering av maximum-likelihood estimationer. (Fox et al (2010)).
Enkelt uttryckt så minimerar en M-estimator en funktion av residualerna. Estimatorns robusthet bestäms av valet av denna funktion. Vid t ex linjär regression väljs denna funktion som summan av residualernas kvadrater.
Genom att välja lämplig funktion kan M-estimatorer vara robusta mot fördelningar med tunga svansar (tex Cauchy) och mot outliers (i y-led). (Andersen (2008)).

Följande teoriavsnitt är hämtat från Fox et al (2010).
Låt oss titta på den linjära modellen där

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i$$

gäller för den i:te av n oberoende observationer. Om modellen antas stämma får vi att $E(y|x) = x_i'\beta$. Givet en estimator $\mathbf{b}$ för β fås den anpassade modellen

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} + e_i = x_i'\mathbf{b}$$

och residualerna fås som

$$e_i = y_i - \hat{y}_i = y_i - x_i'\mathbf{b}$$

För att bestämma $\mathbf{b}$ minimeras en *målfunktion* över alla $\mathbf{b}$,

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x_i'b) \tag{1}$$

där funktionen ρ bestämmer varje residuals bidrag till *målfunktionen*.
ρ ska ha följande egenskaper:

- icke-negativ, $\rho(e) \geq 0$
- lika med noll då argumentet är 0, $\rho(0) = 0$
- symmetrisk, $\rho(e) = \rho(-e)$
- monoton i $|e_i|$, $\rho(e_i) \geq \rho(e_j)$ då $|e_i| > |e_j|$

3.2.1 Beräkning av M-estimation

(Fox et al (2010)).
Det minsta värde på ekvation (1) beräknas genom differentiering med avseende på $\mathbf{b}$, och genom att sätta de partiella derivatorna till 0:

$$0 = \frac{\partial}{\partial b} \sum_{i=1}^{n} \rho(y_i - x_i'b) \tag{2}$$

$$= \sum_{i=1}^{n} \psi(y_i - x_i'b)x_i' \tag{3}$$

där ψ är en kurva definierad som derivatan av ρ med avseende på dess argument.
Definiera *viktfunktionen* som $\omega_i = \omega(e_i) = \psi(e_i)/e_i$. Då kan (3) skrivas som

$$\sum_{i=1}^{n} \omega_i \cdot (y_i - x_i'b)x_i' = 0$$

Lösningen till dessa ekvationer är ekvivalent med ett viktat minstakvadrat problem (minimera $\sum \omega_i^2 e_i^2$). Viktningarna beror dock på residualerna, residualerna beror på de skattade koefficienterna och de skattade koefficienterna beror på viktningarna. Därför krävs en iterativ lösning (IRLS).

1. Välj start estimat $\mathbf{b}^{(0)}$, såsom minsta kvadraten estmatet.
2. Beräkna residualerna $e_i^{(t-1)}$ och de tillhörande vikterna $\omega_i^{(t-1)} = \omega[e_i^{(t-i)}]$ från den föregående iterationen.
3. För nya viktade minsta kvadraten estimatorer lös

$$\mathbf{b}^{(t)} = \left[\mathbf{X}'\mathbf{W}^{(t-1)}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{W}^{(t-1)}\mathbf{y}$$

   där $\mathbf{X}$ är modellmatrisen och $\mathbf{W}^{(t-1)} = diag\{w_i^{(t-1)}\}$ är den aktuella vikt matrisen.

Steg 2 och 3 upprepas tills dess att de beräknade koefficienterna konvergerar. Den asymptotiska kovariansmatrisen för $\mathbf{b}$ är:

$$V(\mathbf{b}) = \frac{E(\psi^2)}{[E(\psi^2)]^2}(\mathbf{X}'\mathbf{X})^{-1}$$

3.3 Breakdown point

Breakdown Point är den andel kraftigt avvikande observationer en estimator klarar av innan resultatet blir felaktigt. Ju högre *Breakdown Point* desto mer robust är estimatorn. Den är mer användbar i situationer med få observationer. Robusta estimatorer med endast en parameter har förhållandevis hög Breakdown Point. En hög Breakdown Point är bra att ha om den kommer på köpet, men den högsta Breakdown Point är oftast svår att hitta. (Huber et al. (2009)).

3.4 Cauchyfördelning

Cauchyfördelningen är en kontinuerlig fördelning med täthetsfunktion:

$$f(x; x_o, \gamma) = \frac{1}{\pi\gamma\left[1 + (\frac{x-x_0}{\gamma})^2\right]} = \frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2 + \gamma^2}\right]$$

När $x_0 = 0$ och $\gamma = 1$ fås standard Cauchyfördelningen med täthetsfunktion:

$$f(x; 0, 1) = \frac{1}{\pi\,(1 + x^2)}$$

För en kontinuerlig slumpvariabel definieras väntevärdet som:

$$E\,(X) = \int_{-\infty}^{\infty} x \cdot f_x(x)dx$$

Eftersom integralen i detta fall är oändlig saknas väntevärde (och därmed varians).
(Alm, Britton. (2008)).

3.5 t-fördelning

t-fördelningen är en kontinuerlig fördelning som till utseendet liknar en normalfördelning. Om en slumpvariabel X har en täthetsfunktion:

$$f(x) = C \cdot \left(1 + \frac{x^2}{f}\right)^{-(f+1)/2}$$

sägs X vara t-fördelad med f frihetsgrader, X∼t(f).
C är en konstant och är:

$$C = \frac{1}{\sqrt{f\pi}} \frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)}$$

För f=1 blir tätheten:

$$f(x) = \frac{1}{\pi} \frac{1}{(1+x^2)}$$

dvs t(1) är samma som Cauchy(0,1). Om X∼t(f) så gäller att:

$$E(X) = 0, \text{om} f > 1.$$

$$V(X) = \frac{f}{f-2} \text{om} f > 2$$

I likhet med Cauchyfördelningen så fås att väntevärde och varians ej existerar då f=1. Då f=2 är väntevärdet 0 men varians existerar ej. (Alm et al. (2008)).

3.6 AR-1 serie

En Autoregressiv (AR) modell är en representation av en slumpmässig process. Modellen beskriver tidsvarierande processer där variabeln är linjärt beroende av sina egna tidigare värden. (Box, Jenkins, Reinsel. (1994) (tredje upplagan)).
För en generell AR(p) definieras modellen enligt:

$$X_t = \sum_{i=1}^{p} \rho_i X_{t-i} + \varepsilon_t, \;\; \varepsilon \sim N(0, \tau^2)$$

där $\rho_1 ... \rho_p$ ($\rho_1 \neq 0$ och $\rho_p \neq 0$) är parametrarna i modellen och där ε_t är

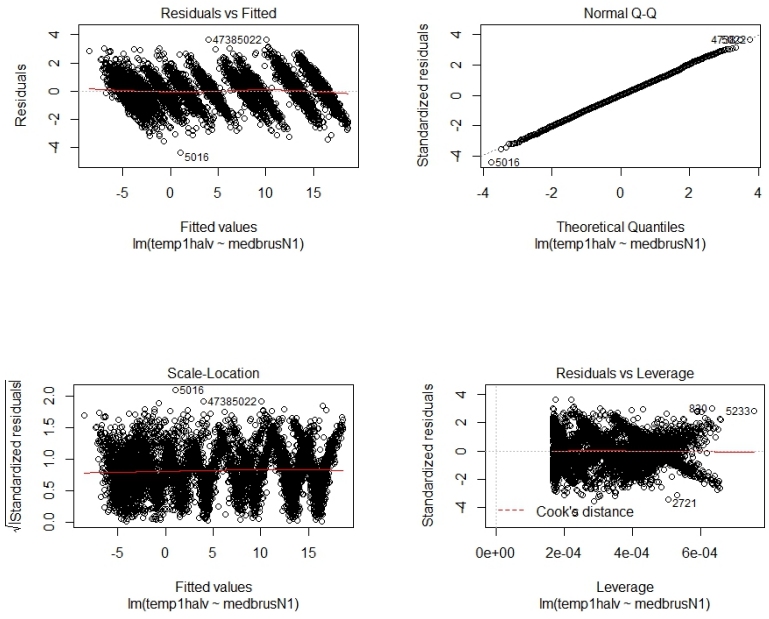
oberoende av varandra. (Fuller. (2008) (Andra upplagan)).
För en AR(1) serie blir modellen följande:

$$X_t = \rho X_{t-1} + \varepsilon_t$$

Ett villkor för att serien ska vara stationär är att $|\rho| < 1$.
(BOX et al (1994)).

4 Analys

Analysen kommer att delas upp i två delar. I den första delen tillförs brus med hjälp av en Cauchyfördelning eller en t-fördelning med 2 frihetsgrader och i den andra delen tillförs brus med hjälp av en AR-1 serie. I analysen jämförs linjär regression (kommandot lm i R) med en robust regression (kommandot rlm i R) som använder sig av en M-estimator. För att kunna jämföra dessa olika modeller görs först en regression av datasetets första halva och därefter en prediktion på andra halvan. Genom att beräkna summan av de kvadratiska avståndet mellan de predikterade värdena och ursprungsvärdena kan modellerna sedan jämföras. Som jämförelse görs här även en linjär regression på datasetet då ett *vitt brus* $\varepsilon \sim N(0, 1)$ tillförts, för att kunna hur de diagnostiska plottarna ser ut utan kraftiga avvikelser.



Figur 2: Diagnostiska plottar med vitt brus

I plotten högst upp till vänster ser vi residualerna plottade mot fitted values. De olika punktsvärmarna i plotten beror på temperaturvariationen mellan årets 12 månader. Trots att det är ett normalfördelat brus som tillförts så visar det sig att residualerna har olika fördelning beroende på vilka fitted values vi tittar på. Detta beror på att bruset är pålagt vågrätt medan regressionen görs lodrät. Eftersom regressionen inte utförts med den *brusiga* variabeln som respons, utan som förklarande variabel så kan vi inte vara säkra på att modellförutsättningarna för linjär regression är uppfyllda ens då bruset är oberoende och normalfördelat med konstant varians.

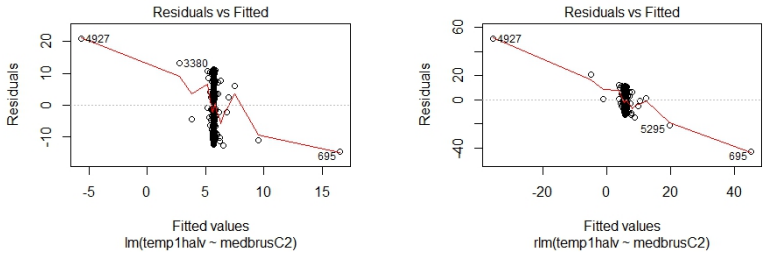
4.1 Brus med hjälp av Cauchyfördelning eller t(2)-fördelning

Eftersom Cauchyfördelningen är en fördelning med tjocka svansar används denna fördelning för att skapa outliers. Genom att generera 6000 Cauchyfördelade värden och addera dessa till 1:a halvan av datasetet fås nu nya värden som ej antas vara normalfördelade. För att få en säker analys upprepas detta 4 gånger. På dessa värden görs både en linjär regression och en robust regression.
Cauchyfördelningen skapar (troligtvis) kraftiga avvikelser från datasetet och därför genereras även 6000 värden från en t-fördelning med 2 frihetsgrader. t-fördelningen liknar en Cauchyfördelning med har inte lika tunga svansar och skapar därför inte lika kraftiga avvikelser. På samma sätt adderas dessa värden till 1:a halvan av datasetet och därefter utförs både linjär regression och robust regression.

4.1.1 Cauchy(0,1)

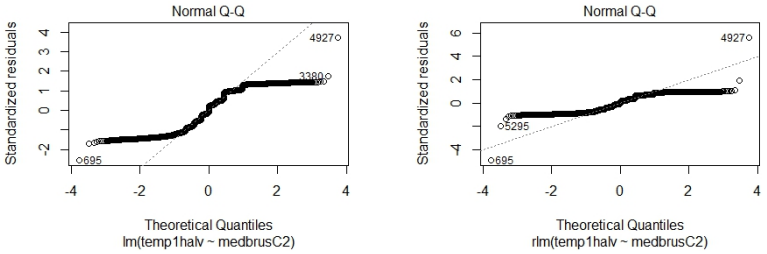
Regression

Till att börja med genereras 6000 värden från Cauchy(0,1) för att användas som brus till 1:a halvan av datasetet.
Nedan visas plottar på residualerna för både linjär regression och robust regression.



Figur 3: Linjär regression till vänster och robust regression till höger.

Båda dess plottar skiljer sig avsevärt från plotten där ett normalfördelat brus tillförts.
För att alla de predikterade värdena ska få plats anpassas skalan, vilket gör att de flesta punkterna bildar en klump i mitten. Vi ser att det finns några kraftigt avvikande observationer vilket är precis vad Cauchyfördelningen ger. Vi ser också att den robusta regressionen ger en större variationsbredd på fitted values, (-20 till 40 för rlm mot -5 till 15 för lm).



Figur 4: Linjär regression till vänster och robust regression till höger.

Plottarna ovan visar (föga förvånande) att vi inte har normalfördelad data. Hade så varit fallet så hade kvantilerna för de olika fördelningarna bildat en rät linje.

Modell	Cauchy(0,1) replikat	Intercept	Koefficient för förklarande variabel
Linjär regression	1	5.6910867	0.0010274
Linjär regression	2	5.6852468	0.0014054
Linjär regression	3	5.6869424	0.0019320
Linjär regression	4	5.6794967	0.0039168
Robust regression	1	5.6208	0.0136
Robust regression	2	5.6610	0.0051
Robust regression	3	5.6257	0.0122
Robust regression	4	5.5620	0.0258

Tabell 1: Tabell över de olika regressionernas intercept och lutningskoefficient

I tabellen ser vi att interceptet för de linjära regressionerna är något högre än för de robusta regressionerna. Vi ser även att koefficienten för den förklarande variabeln är klart högre för de robusta regressionerna. Detta får som följd att variationsbredden blir större för den robusta regressionen än för den linjära, vilket vi också såg i tidigare plottar.

Prediktion

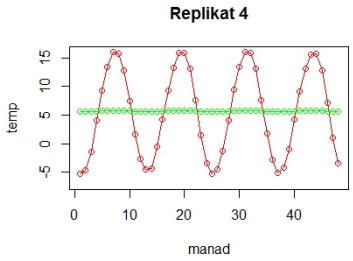
Från de olika regressionsmodellerna (linjär och robust) görs även en prediktion för att ta reda på vilken av modellerna som är att föredra då risken för outliers är stor. Genom att beräkna de kvadratiska avstånden mellan de predikterade värdena och 2:a halvan från ursprungsvärdena kommer den linjära regression och den robusta regressionen att kunna jämföras.

Modell	Cauchy(0,1) replikat	Prediktionsfelens kvadratsumma	Medelvärde av kvadratsumman
Linjär regression	1	351217.4	58.53623
Linjär regression	2	350936.3	58.48939
Linjär regression	3	350585.8	58.43096
Linjär regression	4	349209	58.2015
Totalt för linjär regression			**58.41452**
Robust regression	1	342438	57.073
Robust regression	2	348331.2	58.0552
Robust regression	3	343398.1	57.23301
Robust regression	4	334064.2	55.67737
Totalt för robust regression			**57.00965**

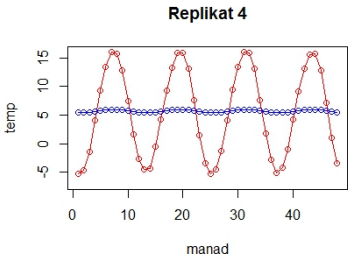
Tabell 2: Tabell över prediktionsfelens kvadratsummor för de olika regressionerna

Från tabellen ovan ser vi att det är den robusta regressionen som ger den minsta totala medelkvadratsumman och därmed anses som den modell som kan förutspå värden bäst. Även om det inte speciellt stor skillnad mellan de olika modellerna, (57.00 för rlm mot 58.41 för lm), så ser vi också att för samtliga 4 olika fall så har den robusta regressionen lägre kvadratsumma för prediktionsfelen. Vi ser också att fall 4 har den lägsta kvadratsumman för respektive modell, vilket kan betyda att det fallet skapade det snällaste bruset.

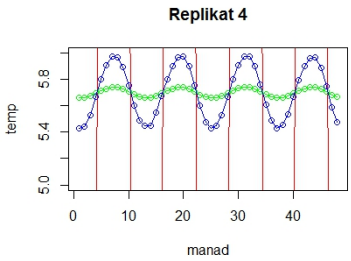
I plottarna nedan visas sanna värden och predikterade värden med tiden som förklarande variabel. För att kunna se något visas endast de 48 första månaderna från 2:a halvan av datasetet och de 48 första predikterade värdena för de olika regressionsmodellerna. I de övre bilderna visas hur originaldatans (röd) och respektive regressions temperaturer varierar mellan årstiderna. Vi ser i plottarna att regressionerna har mycket lägre amplitud än originaldatan, vilket beror på att den förklarande variabeln i regressionmodellerna har så låg inverkan på responsen. I den nedre bilden ser vi regressionsmodellernas toppar (originaldatans toppar syns ej). Vi ser att den robusta regressionen (blå) har högre amplitud än den linjära regressionen (grön).



(a) Originaldata och linjär regression



(b) Originaldata och robust regression



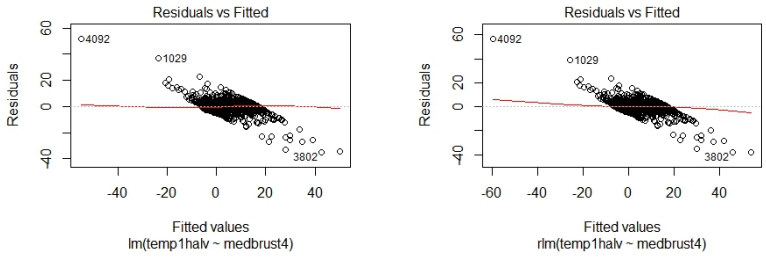
(c) Originaldata och regressionernas prediktioner

Figur 5: Årstidsoscillationer för originaldata (röd), linjär regression (grön) och robust regression (blå).

4.1.2 t(2)-fördelning

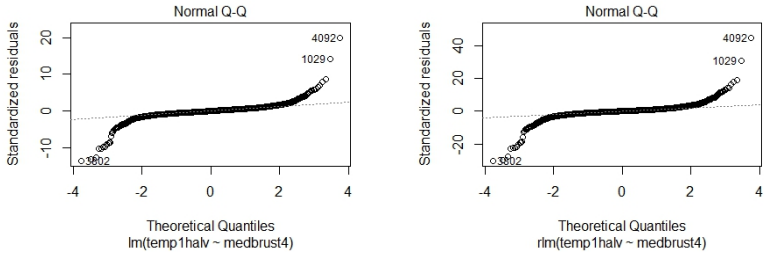
Regression

Som brus genereras 6000 värden från en t-fördelnig med 2 frihetsgrader. Dessa adderas till 1:a halvan av datasetet och linjär och robust regression utförs sedan.
Här visas plottar för både linjär regression och robust regression.



Figur 6: Linjär regression till vänster och robust regression till höger.

Vi ser att plottarna ovan är väldigt lika. Vi ser i båda fallen att plottarna visar på större outliers än i plottarna där normalfördelat brus tillförts. För att alla punkterna ska få plats så väljs en skala som gör att de flesta punkterna bildar en klump i mitten. Plottarna visar också på några avvikande observationer.
I plottarna nedan ser vi att residualerna från regressionerna inte kan antas vara normalfördelade.

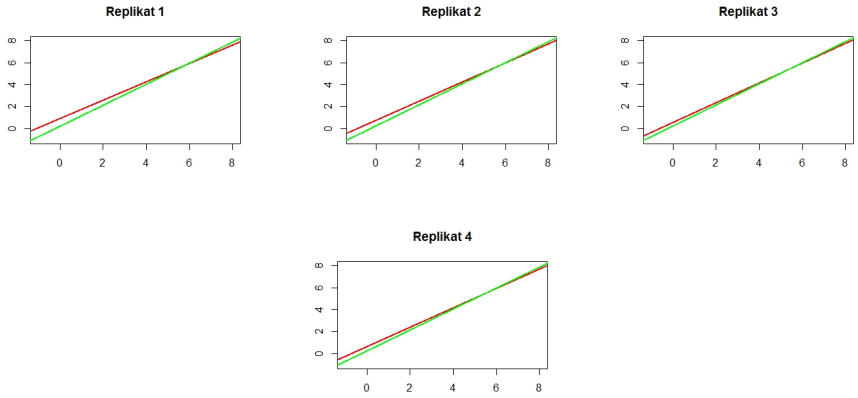


Figur 7: Linjär regression till vänster och robust regression till höger.

Modell	t(2)fördelning replikat	Intercept	Koefficient för förklarande variabel
Linjär regression	1	0.919998	0.836305
Linjär regression	2	0.75901	0.87070
Linjär regression	3	0.56675	0.89997
Linjär regression	4	0.643845	0.882573
Robust regression	1	0.2176	0.9571
Robust regression	2	0.2524	0.9582
Robust regression	3	0.2421	0.9601
Robust regression	4	0.2479	0.9539

Tabell 3: Tabell över de olika regressionernas intercept och lutningskoefficient

I tabellen ser vi att interceptet är mycket högre för de linjära regressionerna än för de robusta regressionerna. Däremot ligger de robusta regressionerna lite högre vad gäller koefficienterna för den förklarande variabeln.



Figur 8: Röd linje motsvarar den linjära regressionen och grön den robusta regressionen.

I plottarna ovan visas grafiska bilder av de olika regressionerna. Vi ser att de linjära regressionslinjerna (röd) är aningen mer vågräta pga lägre koefficient för den förklarande variabeln.

Prediktion

I tabellen nedan visas resultatet av regressionernas prediktioner och deras kvadratsummor.

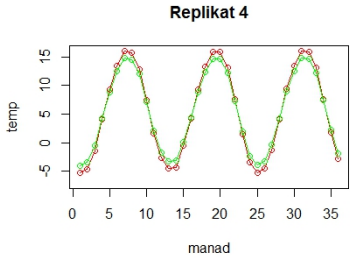
Modell	t(2) fördelning replikat	Prediktionsfelens kvadratsumma	Medelvärde av kvadratsumman
Linjär regression	1	9422.945	1.570491
Linjär regression	2	5898.914	0.9831523
Linjär regression	3	3520.498	0.5867497
Linjär regression	4	4844.421	0.8074035
Totalt för linjär regression			**0.9869491**
Robust regression	1	647.2212	0.1078702
Robust regression	2	618.6365	0.1031061
Robust regression	3	564.0703	0.09401172
Robust regression	4	746.4123	0.124402
Totalt för robust regression			**0.1073475**

Tabell 4: Tabell över prediktionsfelens kvadratsummor för de olika regressionerna

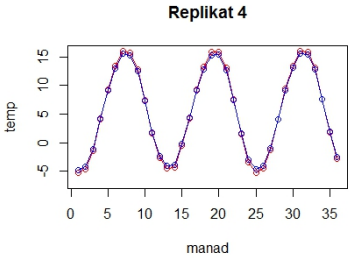
I tabellen ser vi skillnad i regressionernas förmåga att prediktera värden. De robusta regressionerna har lägre kvadratsummor och är det bästa alternativet då ett t(2)-fördelat brus har tillförts. Vi ser också att alla 4 robusta replikat är bättre än de 4 linjära vilket ger en tydlig indikation på att den robusta regressionen är ett bättre alternativ i det här fallet.

På nästa sida visas plottar med sanna värden och predikterade värden där tiden är den förklarande variabeln. För att kunna se något visas endast de 36 första månaderna från 2:a halvan av datasetet och motsvarande 36 första predikterade värdena för de olika regressionsmodellerna.

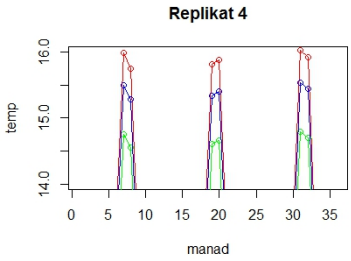
I de övre bilderna ser vi att kurvorna för regressionerna har nästan samma amplitud som originaldatans kurva. Vi ser tydligare i den nedre plotten att den robusta regressionen har en lägre amplitud än originaldata men högre än den linjära regressionen.



(a) Originaldata och linjär regression



(b) Originaldata och robust regression



(c) Originaldata och regressionernas prediktioner

Figur 9: Årstidsoscillationer för originaldata (röd), linjär regression (grön) och robust regression (blå).

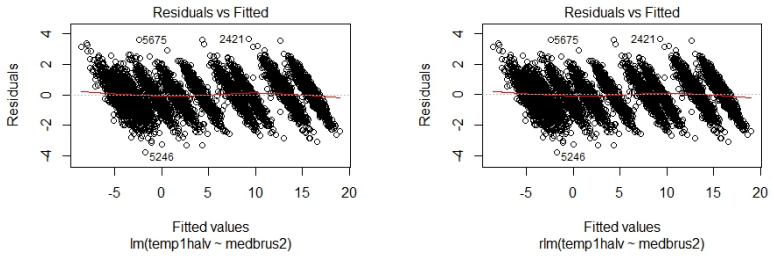
4.2 Brus med hjälp av AR-1 serie

I denna analysdel skapas ett beroende mellan observationerna med hjälp av en AR-1 serie. Bruset skapas genom att generera 6000 värden från en AR-1 serie och sedan addera dessa till 1:a halvan av datasetet. Även här upprepas detta 4 gånger per regression och brus för att få en säkrare analys. För en utförligare analys kommer även parametern ρ att varieras.

4.2.1 $\rho = 0.2$

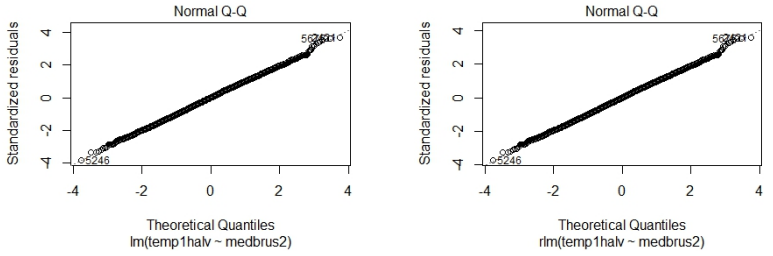
Regression

Till att börja med skapar vi ett brus med $\rho = 0.2$, dvs en ganska låg autokorrelation. Vi får då följande plottar.



Figur 10: Linjär regression till vänster och robust regression till höger.

Vi ser i de övre plottarna att punkterna bildar ett jämt band och inga observationer skiljer sig speciellt mycket från de andra.
I plottarna nedan ser vi att majoriteten av punkterna bildar en rät linje vilket tyder på normalfördelade residualer.



Figur 11: Linjär regression till vänster och robust regression till höger.

Modell	ar(1)-serie, $\rho = 0.2$ replikat	Intercept	Koefficient för förklarande variabel
Linjär regression	1	0.086477	0.982449
Linjär regression	2	0.07441	0.98278
Linjär regression	3	0.12427	0.98129
Linjär regression	4	0.1205	0.9792
Robust regression	1	0.0815	0.9833
Robust regression	2	0.0752	0.9834
Robust regression	3	0.1266	0.9813
Robust regression	4	0.1170	0.9798

Tabell 5: Tabell över de olika regressionernas intercept och lutningskoefficient

I tabellen ser vi att de olika regressionerna ligger väldigt nära varandra för respektive replikat. Vi ser också att interceptet är väldigt lågt (nästan 0) och koefficienten ligger väldigt nära 1 vilket vi också såg i plottarna ovan som visade på ett rakt linjärt samband.

Prediktion

Modell	ar(1)-serie,$\rho = 0.2$ replikat	Prediktionsfelens kvadratsumma	Medelvärde av kvadratsumman
Linjär regression	1	108.5125	0.01808542
Linjär regression	2	106.0075	0.01766792
Linjär regression	3	126.458	0.02107633
Linjär regression	4	152.3816	0.02539694
Totalt för linjär regression			**0.02055665**
Robust regression	1	98.31734	0.01638622
Robust regression	2	97.88864	0.01631477
Robust regression	3	127.0535	0.02117559
Robust regression	4	143.79	0.023965
Totalt för robust regression			**0.0194604**

Tabell 6: Tabell över prediktionsfelens kvadratsummor för de olika regressionerna

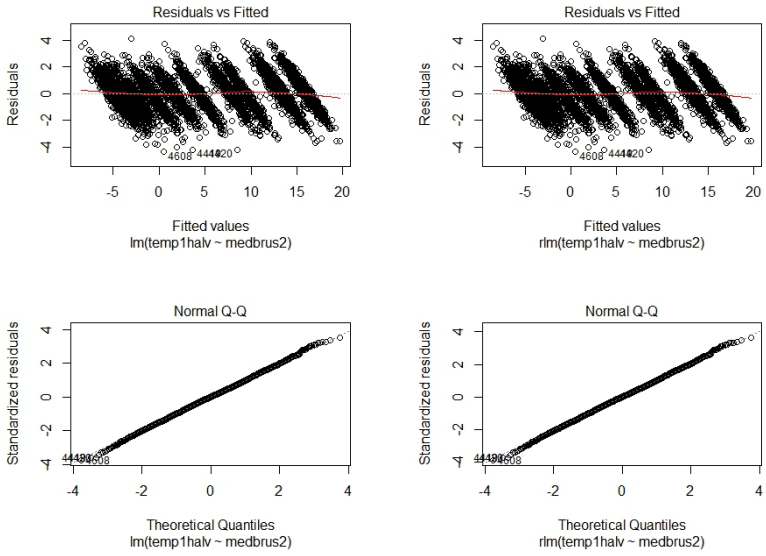
Vi ser att den robusta regressionen har den lägsta kvadratsumman och anses därför som bäst, även om skillnaden är väldigt marginell i detta fall. Vi ser också att för replikat 1,2 och 4 så är den robusta regressionen klart bätttre än den linjära regressionen. För replikat 3 har den linjära regressio-

nen lite lägre kvadratsumma än den robusta regressionen.
När plottar gjordes på de sanna värdena och på regressionernas predikterade värden låg kurvorna väldigt nära varandra. En extrem förstoring visade att den robusta regressionen hade aningen lägre amplitud än de sanna värdena men högre än den linjära regressionen.

4.2.2 $\rho = 0.5$

Regression

Vi ökar autokorrelationen genom att sätta $\rho = 0.5$ och får då följande plottar.



Figur 12: Linjär regression till vänster och robust regression till höger.

I de övre plottarna ser vi att residualerna ligger jämt utspridda kring 0 strecket. Vi ser inga kraftigt avvikande residualer som skulle kunna vara outliers. I de nedre plottarna ser vi att punkterna bildar en rät linje vilket tyder på normalfördelade residualer.

Modell	ar(1)-serie, $\rho = 0.5$ replikat	Intercept	Koefficient för förklarande variabel
Linjär regression	1	0.160250	0.977065
Linjär regression	2	0.135227	0.978839
Linjär regression	3	0.143618	0.978990
Linjär regression	4	0.129430	0.979029
Robust regression	1	0.1572	0.9772
Robust regression	2	0.1347	0.9792
Robust regression	3	0.1420	0.9792
Robust regression	4	0.1412	0.9782

Tabell 7: Tabell över de olika regressionernas intercept och lutningskoefficient

I tabellen ser vi att då $\rho = 0.5$ så fås liknande resultat för linjär och robust regression, dvs regressionslinjen är väldigt lika för modellernas respektive replikat. Även replikat emellan skiljer väldigt lite.

Prediktion

Modell	ar(1)-serie,$\rho = 0.5$ replikat	Prediktionsfelens kvadratsumma	Medelvärde av kvadratsumman
Linjär regression	1	193.2151	0.03220252
Linjär regression	2	160.1795	0.02669658
Linjär regression	3	160.8929	0.02681549
Linjär regression	4	156.2481	0.02604136
Totalt för linjär regression			**0.02793899**
Robust regression	1	190.0283	0.03167139
Robust regression	2	155.2465	0.02587442
Robust regression	3	157.6199	0.02626998
Robust regression	4	170.54	0.02842333
Totalt för robust regression			**0.02805978**

Tabell 8: Tabell över prediktionsfelens kvadratsummor för de olika regressionerna

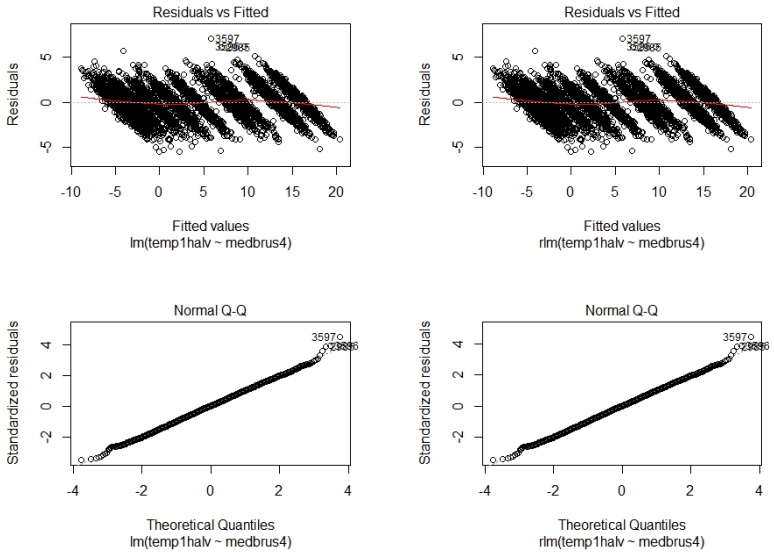
I tabellen ser vi att det är den linjära regressionen som visar bäst resultat (minsta kvadratsumma). Vi ser att för replikat 1,2 och 3 så är resultaten ganska lika men det är den robusta regressionen som har bästa resultat. För replikat 4 är den linjära regressionen klart bättre än den robusta. Så trots att den robusta regressionen var bättre i 3 av 4 replikat så får den linjära regressionen lägst totala kvadratsumma.

Plottar med de sanna värdena och regressionernas predikterade värden gjordes även här. Kurvorna låg väldigt nära varandra där de sanna värdena hade högst amplitud och den linjära regressionen hade lägst.

4.2.3 $\rho = 0.8$

Regression

Vi ökar autokorrelationen ännu mer och sätter ρ till 0.8.



Figur 13: Linjär regression till vänster och robust regression till höger.

I de övre plottarna ser vi att punkterna bildar ett band kring 0 linjen men även att några observationen avviker från mängden. I de nedre plottarna bildas en rät linje centralt men vi ser också att ändarna inte följer en rak linje.

Modell	ar(1)-serie, $\rho = 0.8$ replikat	Intercept	Koefficient för förklarande variabel
Linjär regression	1	0.257480	0.955133
Linjär regression	2	0.258716	0.953690
Linjär regression	3	0.193796	0.953799
Linjär regression	4	0.273125	0.956890
Robust regression	1	0.26752	0.9554
Robust regression	2	0.2535	0.9540
Robust regression	3	0.1903	0.9542
Robust regression	4	0.2766	0.9571

Tabell 9: Tabell över de olika regressionernas intercept och lutningskoefficient

I tabellen ser vi att för tre av de fyra replikaten skiljer värdena för interceptet och koefficienten inte speciellt mycket mellan den linjära regressionen och den robusta regressionen. Vi ser också att replikat 3 i båda regressionerna ger ett lägre intercept men liknande koefficient än de andra tre replikaten.

Prediktion

Modell	ar(1)-serie,$\rho = 0.8$ replikat	Prediktionsfelens kvadratsumma	Medelvärde av kvadratsumman
Linjär regression	1	708.8454	0.1181409
Linjär regression	2	753.9448	0.1256575
Linjär regression	3	766.7373	0.1277896
Linjär regression	4	663.5716	0.1105953
Totalt för linjär regression			**0.1205458**
Robust regression	1	703.6516	0.1172753
Robust regression	2	743.5004	0.1239167
Robust regression	3	754.6571	0.1257762
Robust regression	4	659.7059	0.109951
Totalt för robust regression			**0.1192298**

Tabell 10: Tabell över prediktionsfelens kvadratsummor för de olika regressionerna

Ur tabellen får vi att den robusta regressionen får lägre kvadratsumma och är därmed bättre på att prediktera värden än den linjära regressionen. Det är inte så mycket som skiljer men vi ser också att för samtliga fyra replikat är den robusta regressionen den bästa. Plottar gjordes sedan på de sanna värdena och på regressionernas predikterade värden. I plottarna

låg kurvorna väldigt nära varandra. Vid förstoring visade det sig att den robusta regressionen hade lägre amplitud än de sanna värdena men högre amplitud än den linjära regressionen.

5 Slutsats

Analysen gjordes med hjälp av 5 olika brus där varje del utfördes 4 gånger. Totalt gjordes 20 linjära regressioner och 20 robusta regressioner som sedan jämfördes. Av dessa 20 analyser så visade det sig att kvadratsumman för prediktionsfelen var i 18 av 20 fall lägre för den robusta regressionen. I 4 av 5 olika brus gav robusta regressionen ett bättre resultat än den linjära regressionen. Det var bara då en AR(1)-serien med $\rho = 0.5$ används som brus som den linjära regressionen visade ett bättre resultat totalt för de 4 replikaten. Men i 3 av 4 replikat var även här den robusta regressionen bättre, men då replikat 4 visade på ett klart sämre resultat för den robusta regressionen än den linjära så blev de totala kvadratsummorna lägre för den linjära regressionen.

I 4 av 5 olika brusvarianter så var resultatet mellan regressionerna ganska lika. Det var bara då bruset kom från en t(2)-fördelning som skillnaden på kvadratsummorna mellan regressionerna var stor, (0.11 för robusta regressionen mot 0.99 för den linjära). Här såg vi även den största skillnaden mellan regressionernas respektive intercept och mellan modellernas koefficient för den förklarande variabeln. I årstidsoscillationerna hade båda regressionerna lägre amplitud än originaldata, men den robusta regressionen låg närmast.

I fallet då bruset kom från Cauchy(0,1) och gav kraftiga outliers visade det sig att den robusta regressionen gav en lägre kvadratsumma för prediktionsfelet än den linjära regressionen. Noterbart är dock att interceptet för alla 8 replikat låg mellan 5,56 och 5,69. Koefficienten för den förklarande variabeln låg mellan 0,001-0,004 för de linjära regressionerna och 0,005-0,026 för de robusta regressionerna, dvs den förklarande variabeln hade väldigt låg inverkan på modellen. Som resultat av detta fick de predikterade årstidsoscillationerna en klart lägre amplitud än originaldatan. Den robusta regressionen hade dock lite högre amplitud än den linjära.

Då bruset kom från en AR(1)-serie med $\rho = 0.2$ visade det sig att trots att intercepten och koefficienterna var lika varandra för de olika regressionerna (och deras replikat) blev prediktionsförmågan olika. För 3 av 4 replikat hade den robusta regressionen klart lägre kvadratsummor än motsvarande linjära regression. I replikat 4 var resultaten jämförbara med här var den linjära regressionen lite bättre.

Eftersom den linjära regressionen är lättare att beräkna än den robusta regressionen bör detta tas i beräkning då val av metod görs. Skulle man kunna ana att en störning kommer från en t-fördelning med låg frihetsgrad bör av resultatet att döma en robust regression användas. Men i de övriga fallen är resultaten så lika varandra att en linjär regression ändå är att föredra pga beräkningarna.

I detta projekt användes endast Cauchy(0,1), t(2)-fördelning och en AR(1)-serie med värdena 0.2, 0.5 och 0.8 på ρ. För den intresserade så finns det många andra värden på parametrarna, andra fördelningar och andra sorters brus att laborera med och testa regressionerna på.

6 Referenser

Referenser

[1] ALM, SVEN E., BRITTON, TOM. (Liber 2008). Stokastik.

[2] ANDERSEN, ROBERT. (SAGE 2008). Modern methods for robust regression.

[3] BOX, GEORGE E P., JENKINS, GWILYN M., REINSEL, GREGORY C. (Prentice-Hall 1994) (tredje upplagan). Time series analysis: Forecasting and control.

[4] FOX, JOHN., WEISBERG, SANFORD. (2010). Robust regression in R. (Appendix till An R companion to applied regression (SAGE 2011) (andra upplagan))

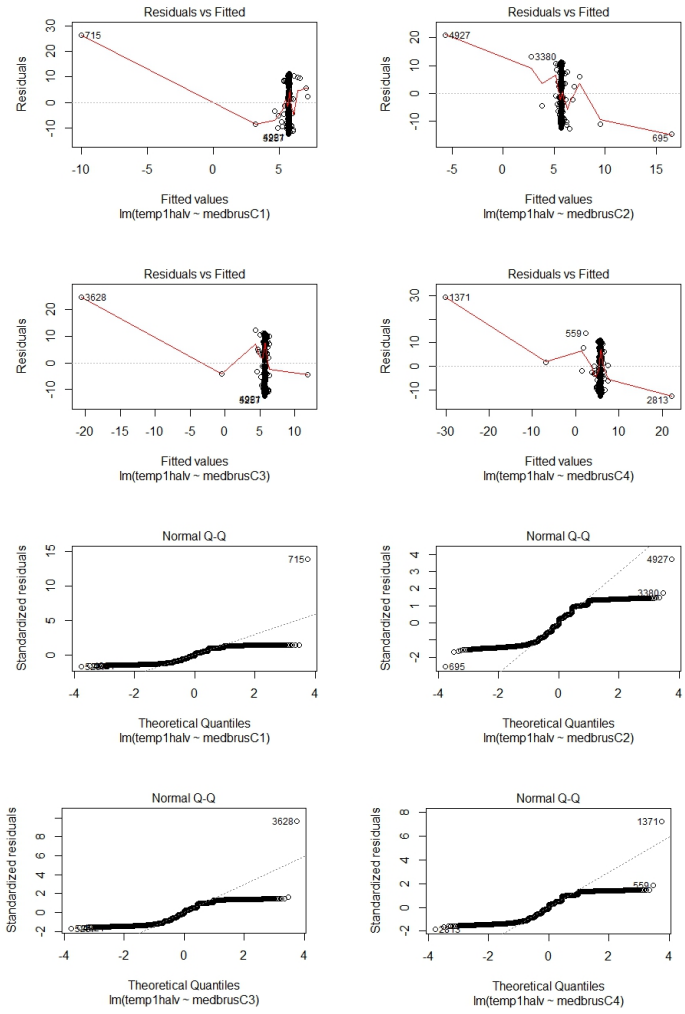
[5] FULLER, WAYNE A. (Wiley 2008) (andra upplagan). Introduction to statistical time series.

[6] HUBER, PETER J., RONCHETTI, ELVEZIO M. (Wiley 2009) (andra upplagan). Robust statistics.

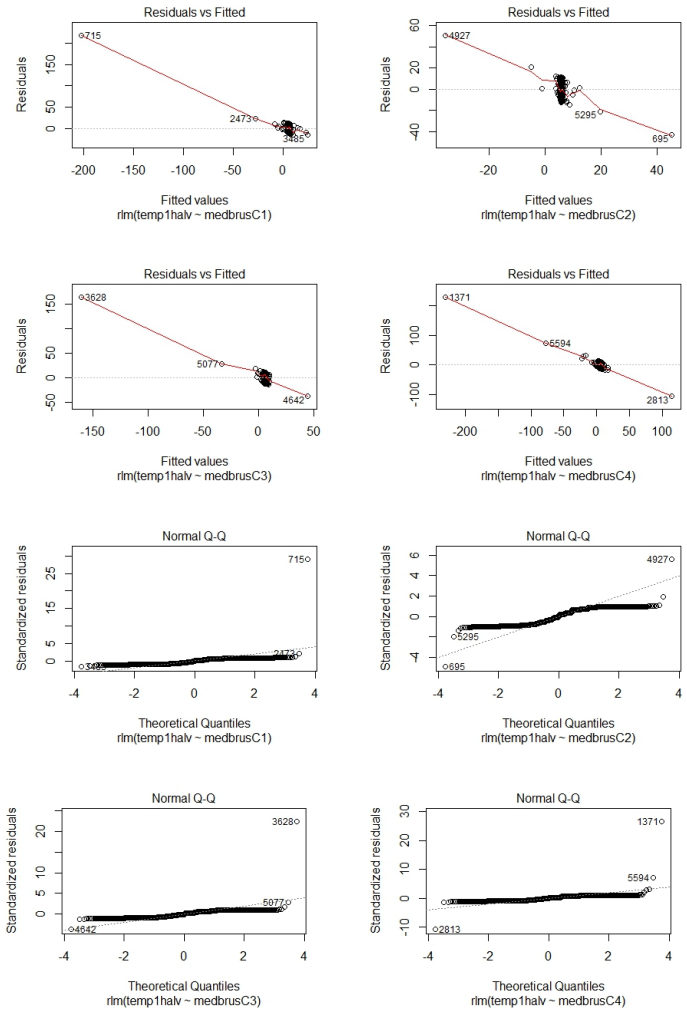
[7] ROUSSEEUW, PETER J., LEROY, ANNICK M. (Wiley 2005). Robust regression and outlier detection.

[8] SUNDBERG, ROLF. (Kompendium Stockholms Universitet 2012). Lineära statistiska modeller

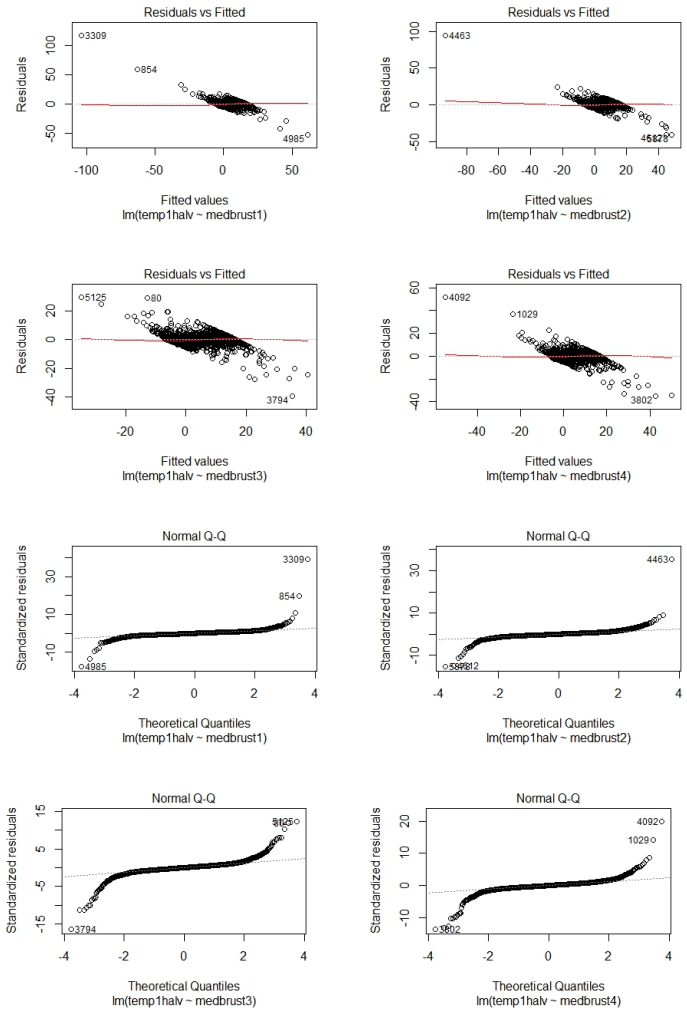
7 Appendix



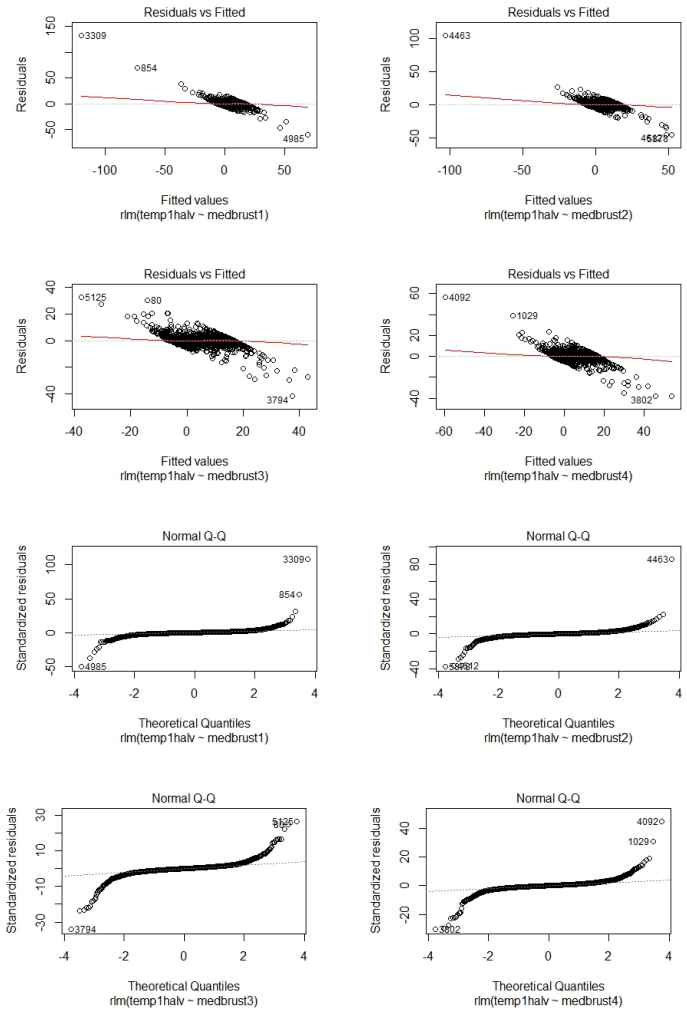
Figur 14: Linjär regression med Cauchy(0,1).



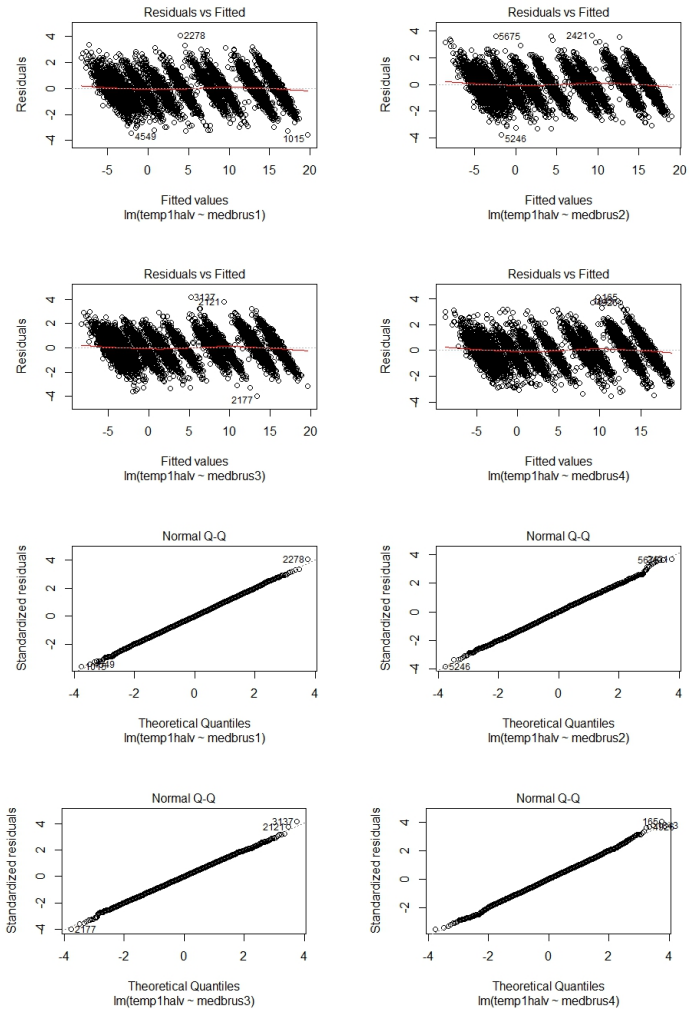
Figur 15: Robust regression med Cauchy(0,1).



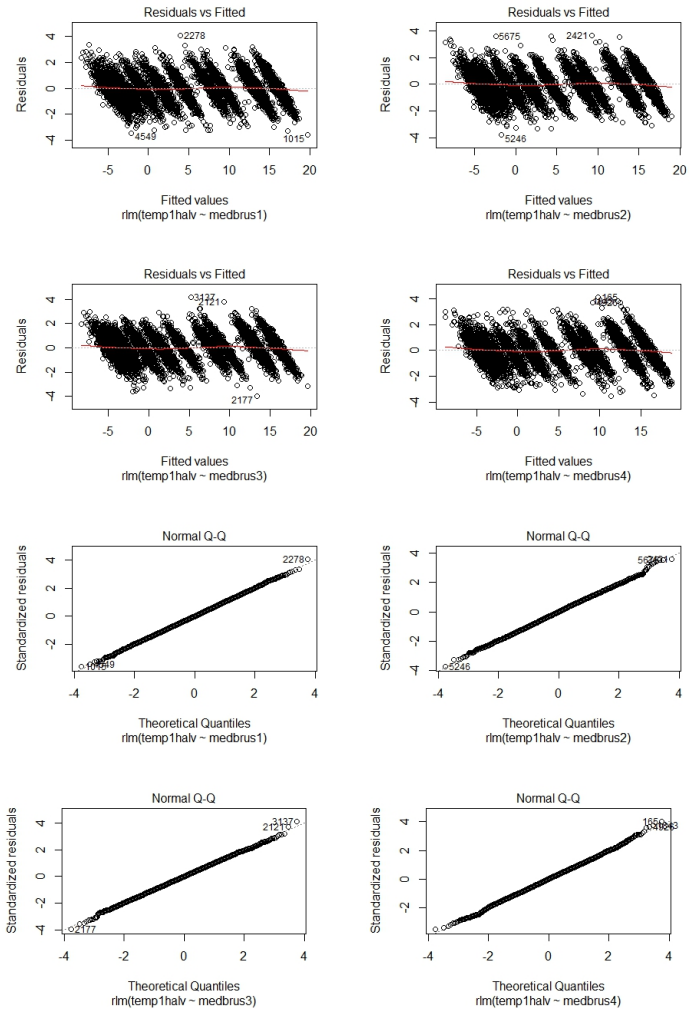
Figur 16: Linjär regression med t(2)-fördelning.



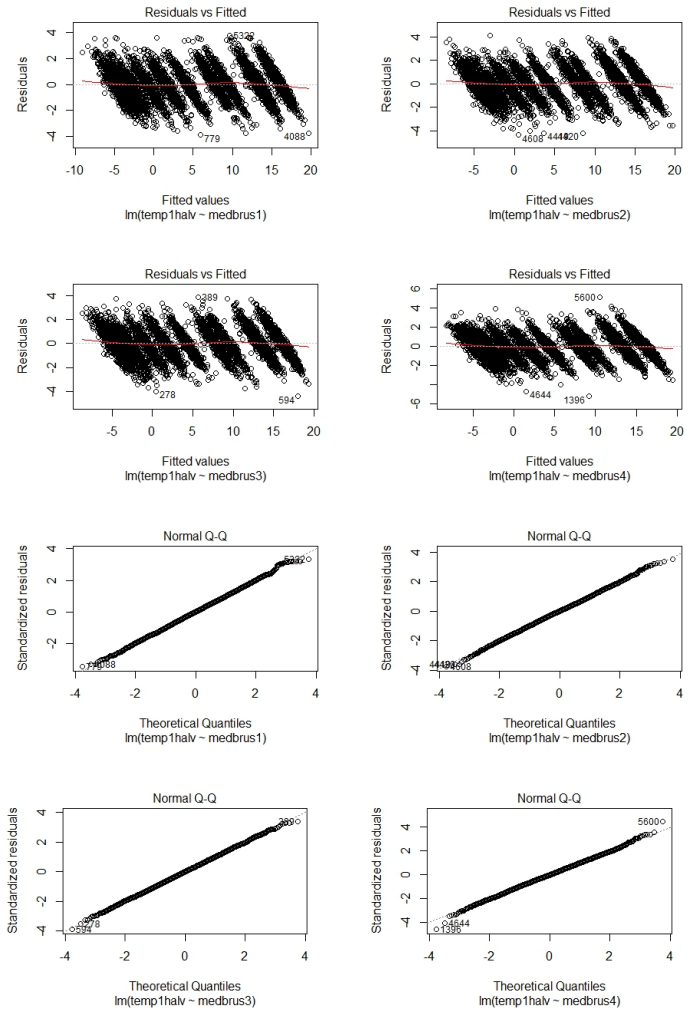
Figur 17: Robust regression med t(2)-fördelning.



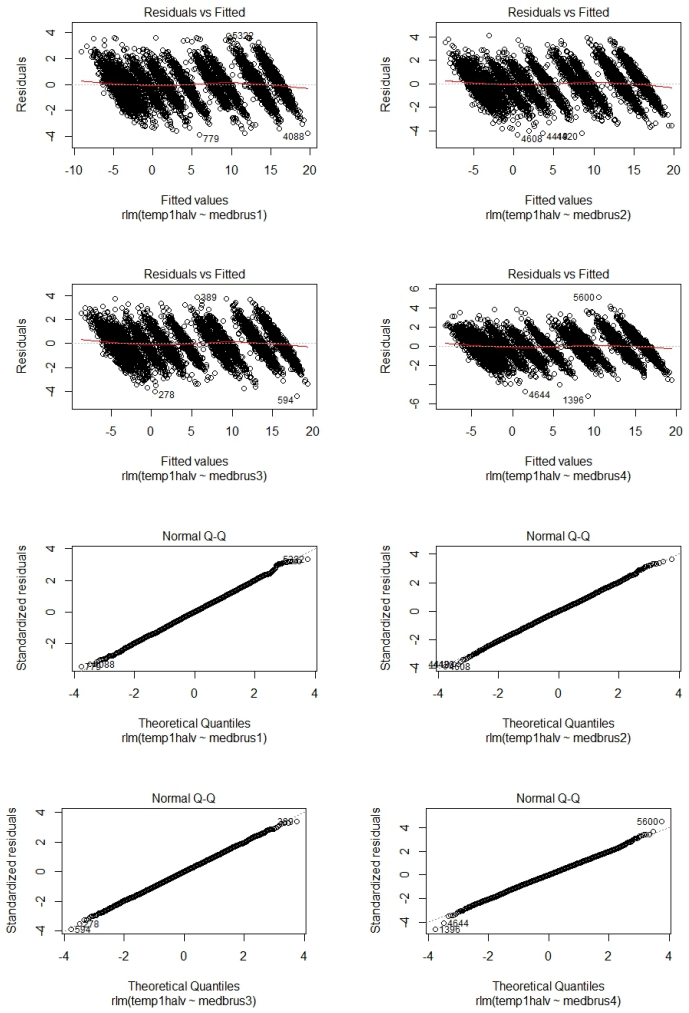
Figur 18: Linjär regression med AR(1), $\rho = 0.2$



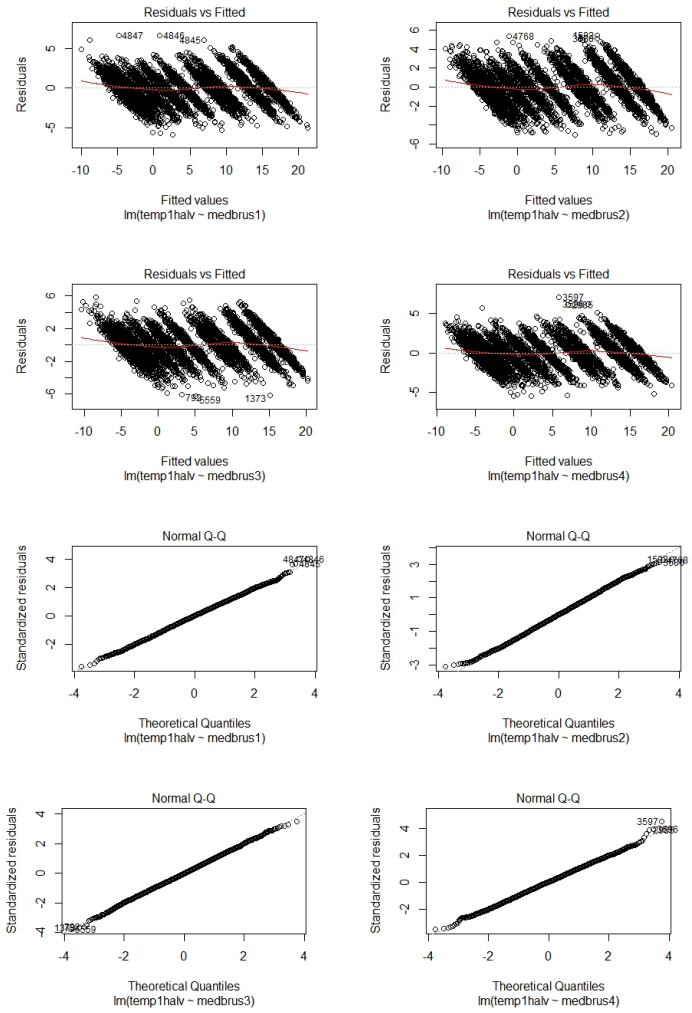
Figur 19: Robust regression med AR(1), $\rho = 0.2$.



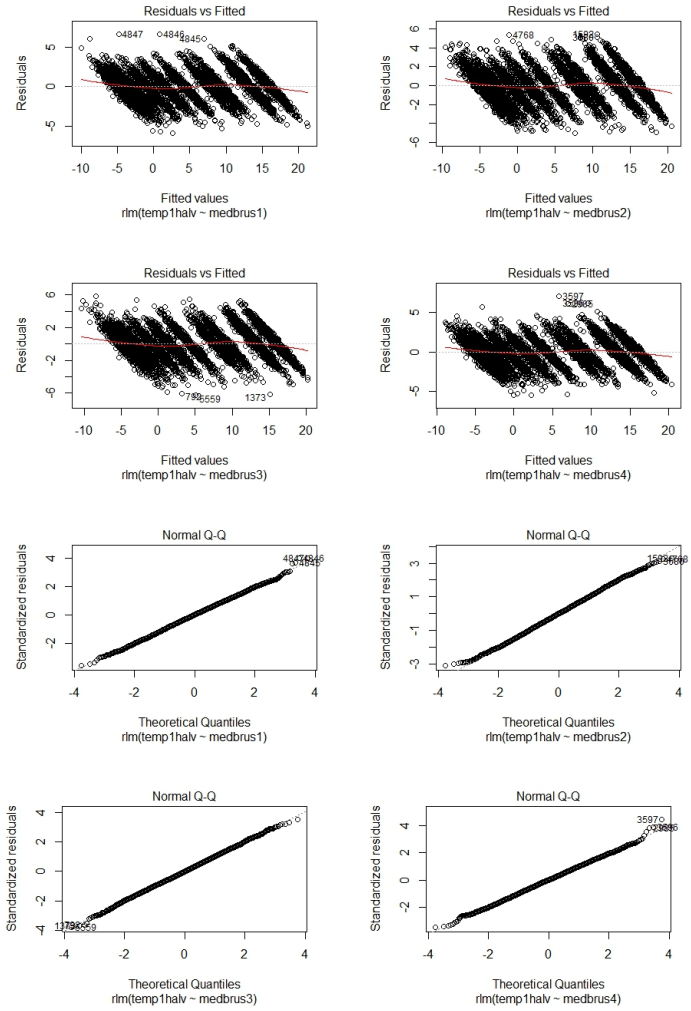
Figur 20: Linjär regression med AR(1), $\rho = 0.5$.



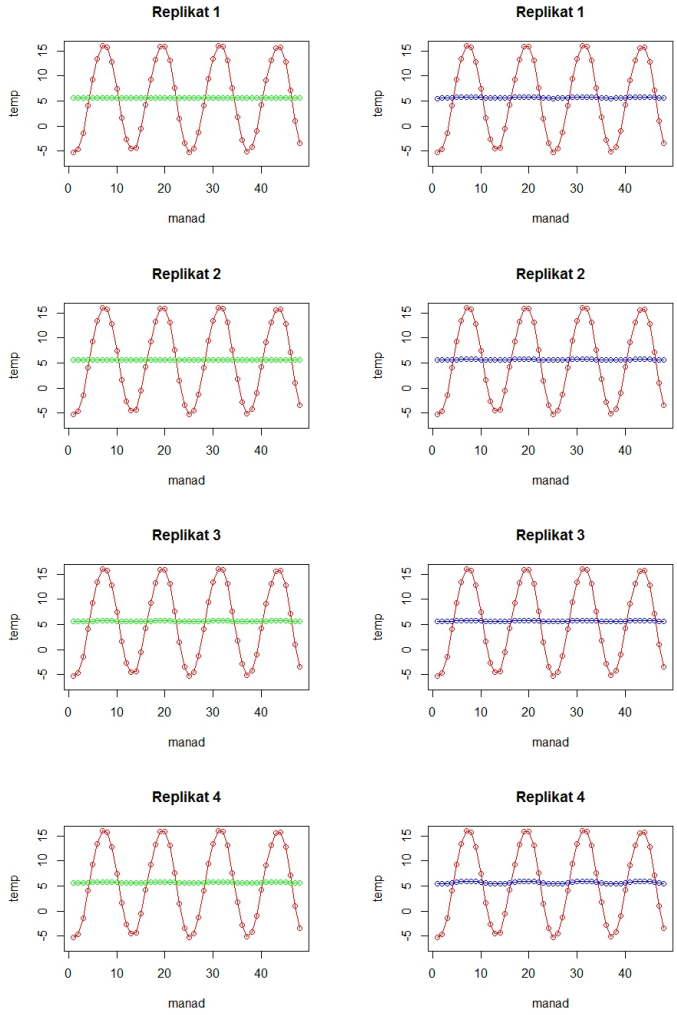
Figur 21: Robust regression med AR(1), $\rho = 0.5$.



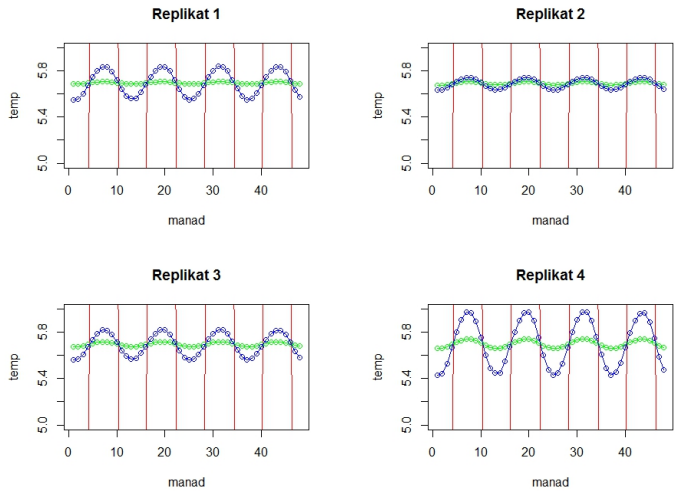
Figur 22: Linjär regression med AR(1), $\rho = 0.8$.



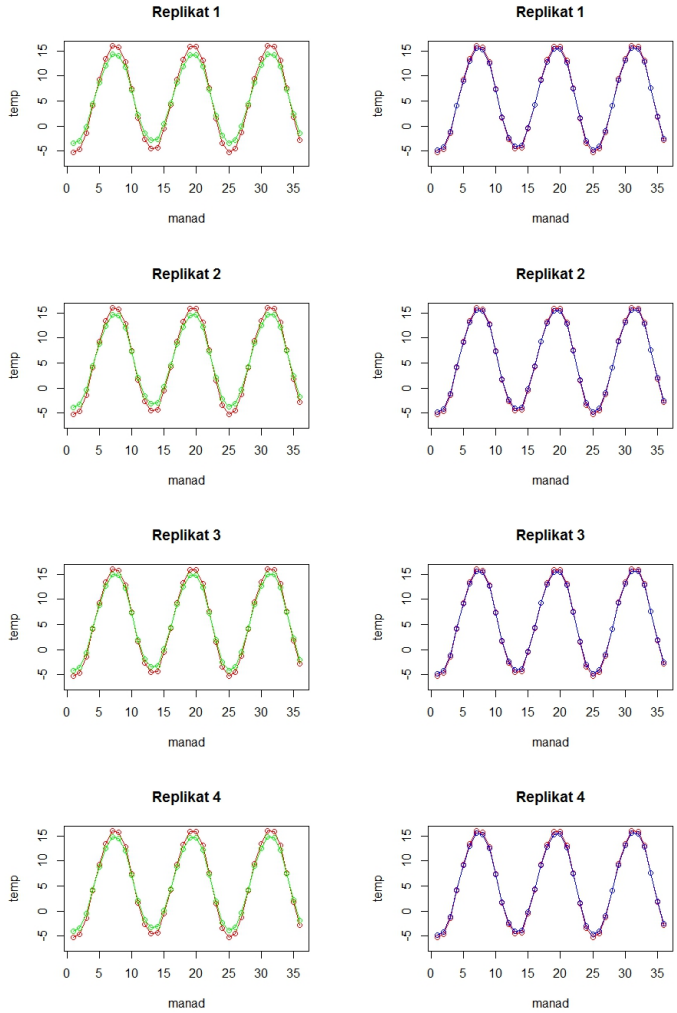
Figur 23: Robust regression med AR(1), $\rho = 0.8$.



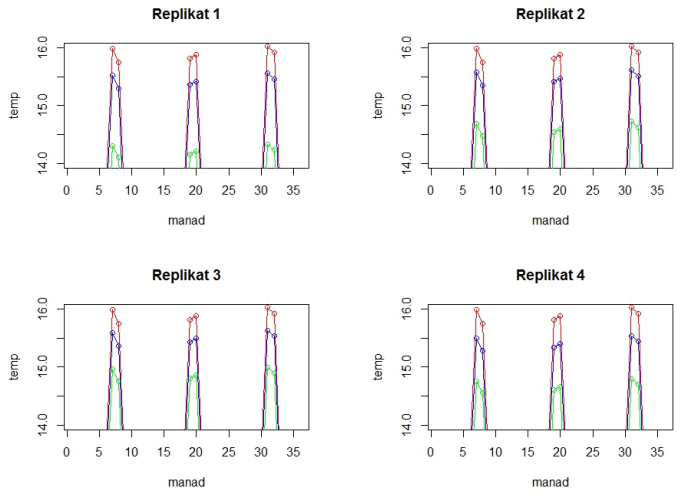
Figur 24: Årstidsoscillationer för originaldata (röd), linjär regression (grön) och robust regression (blå). Brus som Cauchy(0,1)



Figur 25: Årstidsoscillationer för originaldata (röd), linjär regression (grön) och robust regression (blå). Brus som Cauchy(0,1)



Figur 26: Årstidsoscillationer för originaldata (röd), linjär regression (grön) och robust regression (blå). Brus från en t(2)-fördelning



Figur 27: Årstidsoscillationer för originaldata (röd), linjär regression (grön) och robust regression (blå). Brus från en t(2)-fördelning