Stockholms universitet

Statistisk analys av sambandet mellan geometriska parametrar och snurrtal på ett XPI cylinderhuvud

Ekaterina Fetisova

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2010:10
Matematisk statistik
September 2010

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2010:10**,
http://www.math.su.se/matstat

Statistisk analys av sambandet mellan geometriska parametrar och snurrtal på ett XPI cylinderhuvud

Ekaterina Fetisova*

September 2010

Sammanfattning

Detta examensarbete redovisar en statistisk analys av sambandet mellan geometriska parametrar och snurrtal på ett XPI cylinderhuvud. På Scania, där det nya XPI cylinderhuvudet har utvecklats, anser man att kännedom om detta samband är avgörande för vidare utveckling av Scanias lastbilsmotorer. Totalt har 120 cylinderhuvuden och 35 parametrar analyserats. Elimination av 6 parametrar ledde till en väsentlig minskning av graden av multikollinearitet och det möjliggjorde tillämpning av linjära regressionsmetoder. Resultatet av regressionerna har dock visat att det inte går att bestämma en slutgiltig modell som återspeglar sambandet mellan geometriska parametrar och snurrtal på det bästa sättet. Därför har sju möjliga modeller presenterats. Analysen av dessa modeller ger i sin tur en viss beskrivning av vilka parametrar som kan påverka snurrtalet. Det finns dock anledning att tro att några viktiga parametrar har saknats i studien. Genomförandet av ett kontrollerat experiment, exempelvis ett faktorförsök, minskning av antalet parametrar samt undersökning av nya möjliga parametrar kan leda till andra men samtidigt säkrare slutsatser

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: parsenergi@tele2.se. Handledare: Jan-Olov Persson.

Statistical analysis of the relationship between some geometrical parameters and a swirl number of an XPI cylinder head

Abstract

This thesis describes a statistical analysis of the relationship between some geometrical parameters and a swirl number of an XPI cylinder head. The new XPI cylinder head has been developed by Scania and the knowledge of this relationship is considered there as crucial for the further development of Scania's motors for trucks. Altogether 120 cylinder heads and 35 parameters have been included in the analysis. Deletion of 6 parameters led to an essential reduction of multicollinearity, which makes it possible to apply linear regressions methods. However the results of the regressions have showed that it is probably impossible to define only one final model, which represents the relationships in the best way. Therefore seven possible final models have been defined. The analysis of these modeles gives a certain description of the parameters which kan affect the swirl number, though there is a reason to suppose that some important parameters for the swirl number has not been included in the analysis. Conducting a planned experiment, for instance the 2^p design, reducing the total number of original parameters but at the same time incorporating possible alternative parameters can lead to different but more certain conclusions.

Förord

Det här analysen är genomförd på uppdrag av Scania och är utfört vid Matematiska institutionen vid Stockholms universitet samt på Scania i Södertälje. Jag vill tacka min handledare på Matematiska institutionen, Jan-Olov Persson, för hans stöd och värdefulla råd inom statistik samt Daniel Granquist (student, KTH) och hans handledare Stefan Elfvin (Scania) för deras samarbete.

Observera att viss information har tagits bort ur rapporten av sekretesskäl på begäran av Scania. Detta betyder att vissa plottar saknar skala samt att alla uppgifter om variablernas spridning har eliminerats.

Innehåll

1 Inledning

1.1 Bakgrund

Den skärpta emissionslagsstiftningen för förbränningsmotorer leder till en förändrad kravsättning på förbränningssystemet i Scanias motorer. Scania har för emissionsnivå Euro 5 valt att följa två olika strategier, EGR (Exhaust Gas Recirculation) och SCR (Selective Catalyc Reduction) för att uppnå dessa krav. För att få ett förutsägbart förbränningsförlopp är en viktig konstruktionsparameter hur mediet i förbränningsrummet beter sig vid insugs- och kompressiontakt. Detta beteende kan beskrivas genom det snurrtal som cylinderhuvudet har. För att säkerställa att kravsättningen på cylinderhuvudet är korrekt och att funktionen är robust behöver sambandet mellan geometriska parametrar och snurrtal undersökas.

1.2 Syfte

På tidigare cylinderhuvuden som Scania har tagit fram så har man sett ett klart samband mellan snurrtal och vissa geometriska parametrar men med det nya XPI cylinderhuvudet så kan man inte se att dessa tidigare kända parametrar själva påverkar snurren. Detta kandidatexamensarbete utgör den statistiska delen av ett annat kandidatexamensarbete som genomförs av Daniel Granquist (KTH, maskinteknik) och det har till sitt syfte att genomföra en statistisk analys om hur geometriska parametrar inverkar på XPI cylinderhuvudets snurrtal. Med andra ord är syftet att bygga en statistisk modell som beskriver det ovannämda sambandet på ett enkelt men tillfredställande sätt. Dessutom strävar man efter att den slutgiltiga modellen ska ha en tillräckligt hög prediktionsförmåga för att kunna användas för prediktion av snurrtal.

1.3 Analysmetod

Huvudmetoden som skall användas i den här analysen är `multipel linjär regression`. Denna metod består i anpassning av modeller för en beroende variabel som en funktion av några förklarande variabler. Resultat av anpassningen blir dock tillförlitliga om förklarande variabler inte är kraftigt korrelerade. För att uppfylla kravet skall variabelelimination genomföras. Sättet att eliminera variabler förväntas att leda till en minimal informationförlust.

2 Teori

2.1 Multipel Linjär Regression

2.1.1 Definition

Antag att det finns p *förklarande variabler* X_1, X_2,..., X_p som är relaterade till en responsvariabel Y samt att data kommer från ett slumpmässigt stickprov av storlek n $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, $i = 1, 2, \ldots, n$. Modellen för multipel linjär regression i termer av observationer definieras då som ([4] sid 220):

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad i = 1, 2 \ldots, n \qquad (2.1)$$

där:
(1) $\epsilon_i \sim N(0, \sigma^2)$ är de icke-observerade slumpmässiga felen, som är oberoende
(2) $E[Y_i] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ samt $Var(Y_i) = \sigma^2$
(3) talen x_{ij} förutsätts vara kända utan slumpmässiga fel.

De okända parametrarna α, β_1, β_2, ..., β_p skattas med hjälp av *minsta-kvadrat-metoden* (MK-metoden). Om $\mathbf{X}^t\mathbf{X}$ är icke-singulär så är MK-skattningen

$$\hat{\theta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

där $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p)^t$ och

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \qquad .$$

I allmänhet kan vilken modell som helst som är linjär i parametrarna studeras med hjälp av multipel regressionsanalys.

2.1.2 Standardiserade regressionskoefficienter

Relationen mellan de ursprungliga regressions- och standardiserade koefficienterna, β_j respektive b_j, ges av ([5] sid 170):

$$\hat{\beta}_j = \hat{b}_j \frac{s_y}{s_j}, \qquad j = 1, 2, \ldots, p$$

där

$$s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1}$$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Med standardiserade koefficienter får modellen (2.1) följande utseende

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + \ldots + b_p z_{ip} + \epsilon_i, \qquad i = 1.2, \ldots, n \tag{2.2}$$

där

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p. \tag{2.3}$$

$$y_i^* = \frac{y_i - \bar{y}}{s_y}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p. \tag{2.4}$$

På så sätt har alla standardiserade förklarande variabler och den standardiserade responsvariabeln stickprovsmedelvärde, som är lika med noll, och stickprovsvarians, som är lika med 1.

Vidare bör följande samband nämnas

$$\mathbf{Z}^t\mathbf{Z} = (n-1)\begin{bmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1p} \\ r_{21} & 1 & r_{23} & \ldots & r_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \ldots & 1 \end{bmatrix} \tag{2.5}$$

där $\mathbf{Z}$ är en $n \times p$ matris av z_{ij} (2.3) samt den angivna matrisen i högra ledet är korrelationsmatrisen med

$$r_{ij} = \frac{\sum_{u=1}^{n}(x_{ui} - \bar{x}_i)(x_{uj} - \bar{x}_j)}{\sqrt{\sum_{u=1}^{n}(x_{ui} - \bar{x}_i)^2 \sum_{u=1}^{n}(x_{uj} - \bar{x}_j)^2}} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} \tag{2.6}$$

som är korrelationen mellan förklarande variabler x_i och x_j.

Vid standardisering av variabler på det här sättet undviker man numeriska problem, som kan uppstå om de ursprungliga variablerna skiljer sig betydligt i magnitud. Standardiserade koefficienter kan användas som ett mått av den relativa betydelsen av förklarande variabler, som ingår i modellen. Variabel med den största standardiserade koefficienten betraktas då som den viktigaste. Men det är viktigt att komma ihåg att $\hat{b}_j$ är påverkade av variablernas variationsvidd. Detta kan leda till andra slutsatser om variablernas betydelse vid analys av ett annat datamaterial, där variablerna varierar annorlunda. ([5] sid 170). Vidare blir det inte alltid lätt att tolka standardiserade koefficienter.

2.1.3 Added Variabel Plot

I enkel regression kan relationen mellan responsvariabeln Y och den förklarande variabeln X analyseras med hjälp av scatterplot. I multipel regression blir situationen mer komplicerad på grund av relationer mellan själva

förklarande variabler och deras simultana inverkan på responsen. *Added variabel plottar* kan hjälpa oss att se effekten av X_k på Y, givet att det finns andra *p -1* förklarande variabler i modellen (2.1) ($1 \leq k \leq p$). Sådana plottar erhålles så här [7]:

1. Regression av Y på alla X förutom X_k ger residualerna $\hat{\epsilon}_Y(X_k)$ som representerar den del av Y som inte förklarades av alla X förutom X_k.

2. Regression av X_k på alla andra X ger residualerna $\hat{\epsilon}_k$ som representerar den del av X_k som inte förklarades av de andra X.

3. Plotta $\hat{\epsilon}_Y(X_k)$ mot $\hat{\epsilon}_k$. Ett starkt linjärt samband mellan de plottade residualerna motsvarar ett starkt justerat samband mellan X_k och Y. Om plotten inte visar någon stark trend då är den justerade relationen svag. Generellt kan den typen av plot tolkas på samma sätt som scatterplot i en enkel regression.

Korrelationen mellan $\hat{\epsilon}_Y(X_k)$ och $\hat{\epsilon}_k$ kallas för *partiell korrelation* mellan Y och X_k justerad för övriga X och den ska betecknas i det här arbetet med r^* för att skilja den från korrelationen mellan Y och X i en enkel regression, som här betecknas med r.

2.1.4 Partiell Residualplot

För att se om den förklarande variabeln X_k behöver transformeras gör man *partiell residualplot*, i vilken $\hat{\epsilon}_i + \hat{\beta}_k x_{ik}$ plottas mot x_{ik}, där $\hat{\epsilon}_i = y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}$ är residualerna från modellen (2.1). Om plotten visar något krökt samband bör variabeln transformeras ([7]).

2.1.5 Anpassningsmått: R^2 och Justerad R^2

Förklaringsgraden, R^2, talar om hur stor del av den totala variationen som förklaras av variationen i de förklarande variablerna ([6] sid 68)

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R^2 ökar alltid om en ny x-variabel tillförs till en modell även om den variabeln inte förklarar nämnvärt.

Justerad R^2 (eng. *adjusted* R^2) definieras i sin tur som

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1 - R^2) \quad .$$

Till skillnad från R^2 behöver R^2_{adj} inte öka om en ny x-variabel tillförs till en modell vilket beror på justeringen för frihetsgrader, $(n-1)/(n-p-1)$.

2.2 Jämförelse av regressionsmodeller

Förutom *kvantitativa variabler* eller kontinuerliga variabler kan även *kvalitativa* variabler eller kategorivariabler användas i multipel regression som förklarande variabler. I detta stycke kommer alla definitioner från [7]. Betrakta en modell som innehåller en *kvantitativ* variabel X och en *kvalitativ* variabel Q med 3 nivåer som blir kodad med hjälp av indikator variabler på följande sätt:

		Q_1	Q_2	Q_3
nivåer	1	0	0	0
	2	0	1	0
	3	0	0	1

Möjliga modeller som beskriver effekten av Q på relationen mellan X och Y ges av

$$Y = \alpha + \beta X + \gamma_1 Q_2 + \gamma_2 Q_3 + \rho_1 X Q_2 + \rho_1 X Q_3 + \epsilon \tag{2.7}$$

$$Y = \alpha + \beta X + \gamma_1 Q_2 + \gamma_2 Q_3 + \epsilon \tag{2.8}$$

$$Y = \alpha + \beta X + \epsilon \tag{2.9}$$

Modellen (2.7), den *allmänna* modellen, förutsätter 3 regressionslinjer med olika intercept och olika lutningskoefficienter.

Modellen (2.8), modell *Parallella Regressioner*, förutsätter 3 regressionslinjer med olika intercept men med samma lutningskoefficient, dvs de är parallella.

Modellen (2.9), model *Sammanfallande Regressioner*, förutsätter en och samma linje för varje nivå av Q.

Med andra ord förutsätter modellen (2.7) och (2.8) olika regressionsmodeller för varje nivå av Q, medan modellen (2.9) förutsätter en och samma modell.

Jämförelse av modellerna (2.8) och (2.9) med modellen (2.7) innebär att följande nollhypoteser testas:

$$H_0 : \gamma_1 = \gamma_2 = 0$$
$$H_0 : \gamma_1 = \gamma_2 = \rho_1 = \rho_2 = 0$$

Alternativhypotesen blir i sin tur

$$H_A : \text{minst en av} \quad \rho_i \quad \text{och/eller} \quad \gamma_j \quad \text{är skild från noll,} \quad i, j = 1, 2.$$

Dessa två test kan utföras genom att använda F-statistika

$$F = \frac{(SSE_{red} - SSE)/q}{SSE/(n - p - 1)} \tag{2.10}$$

där SSE är residualkvadratsumman i modellen (2.7) och SSE_{red} är residualkvadratsumman i en av de reducerade modellerna (2.8) eller (2.9), q betecknar antal parametrar som sätts lika med noll i H_0. Om den reducerade modellen avviker kraftigt från den fullständiga modellen (2.7) blir F-statistikan stor jämfört med $F_\alpha(q, n-p-1)$.

Att testa modell (2.9) mot modell (2.8) brukar kallas för *analys av kovarians* ([5] sid 184).

2.3 Multikollinearitet

2.3.1 Definition

Förklarande variabler X_1, X_2, ..., X_p är linjärt beroende om det finns konstant c_0 och konstanter c_1, c_2, ..., c_p, ej alla lika med noll, sådana att ([5] sid 288):

$$c_1X_1 + c_2X_2 + \ldots + c_pX_p = c_0 \tag{2.11}$$

Om (2.11) gäller approximativt för någon uppsättning av förklarande variabler, då säger man att nästan linjärt beroende råder i $\mathbf{X}^t\mathbf{X}$ och nästan kollinearitet existerar. Nästan multikollinearitet innebär att det finns några uppsättningar av förklarande variabler som approximativt uppfyller (2.11).

Multikollinearitet har bland annat följande effekter:

1. Det leder till stora varianser för skattade koefficienter, vilket medför långa konfidensintervall för β_j och som följd av detta ger ett dåligt grepp om vad β_j egentligen är.

2. Det resulterar ofta i skattade koefficienter som är icke-signifikanta, fastän motsvarande variabler kan ha betydelse för responsen, eller koefficienter som har felaktiga tecken eller magnitud. Detta är ett problematiskt resultat särskilt när det är av intresse att bestämma struktur av relationer mellan responsen och förklarande variabler.

2.3.2 Diagnostik av multikollinearitet

Det finns olika diagnostikmått för att upptäcka multikollinearitet. Här diskuteras några av dem.

`1. Undersökning av korrelationsmatrisen`

Höga parvisa korrelationer r_{ij} (2.6) indikerar existens av parvis multikollinearitet.

2. VIF, Variance Inflation Factors

I fallet då $p > 2$ kan det visas att variansen av den $j$te koefficienten är ([7] sid 198)

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \left(\frac{1}{S_{jj}} \right) \qquad j = 1, 2, \ldots, p \tag{2.12}$$

där R_j^2 är förklaringsgraden av regression av X_j på de andra $p-1$ förklarande variablerna. Om det finns ett starkt beroende mellan X_j och någon uppsättning av de andra $p-1$ variablerna, då ska värdet på R_j^2 vara nära 1. Detta medför att $\mathrm{Var}(\beta_j)$ blir stora. Faktorn $\frac{1}{1-R_j^2}$ kallas för $j$te *variance inflation factor*, eller VIF_j och den representerar ökningen i varians på grund av korrelationen mellan förklarande variabler, dvs. kollinearitet. VIF_j som är större än 5 eller 10 är en indikation på att motsvarande koefficient är dåligt skattad på grund av multikollinearitet ([5] sid 300).

3. Analys av egensystem av $\mathbf{X}^t\mathbf{X}$

Analys av egensystem består i analysen av egenvärden och egenvektorer till matrisen $\mathbf{X}^t\mathbf{X}$, där $\mathbf{X}$ är en matris med n rader och p' kolumner [1]. Egenvärden, λ_1,λ_2,..., $\lambda_{p'}$, som är lika med noll indikerar exakt linjärt beroende mellan kolumner i $\mathbf{X}$, medan egenvärde som är nära noll indikerar nästan linjärt beroende. Om det finns ett eller flera nästan linjärt beroende finns det ett eller flera egenvärden som är små. Egenvektor som hör ihop med ett litet egenvärde kan användas för att bestämma variabler som bidrar till multikollinearitet - elementen i denna egenvektor som är relativt stora indikerar variabler som bidrar mest ([4] sid 281). Om måttenheterna av kolumnerna i $\mathbf{X}$ är godtyckliga beräknas egenvärde till korrelationsmatrisen (se sambandet 2.5).

Vidare definieras *konditionstal* k som

$$k = \frac{\lambda_{max}}{\lambda_{min}} \qquad . \tag{2.13}$$

Detta är ett mått på spridning hos egenvärden till $\mathbf{X}^t\mathbf{X}$. Allmänt om k är mindre än 100 finns det inte allvarliga problem med multikollinearitеten. Konditionstal som är mellan 100 och 1000 medför måttlig till stark multikollinearitet, och k som är större än 1000 indikerar grov multikollinearitet ([5] sid 301).

[1] p' står för antal parametrar: $p' = p$ för modell utan intercept och $p' = p+1$ för modell med intercept

2.4 Prediktion och modellval

Här diskuteras två kriterier för modellval som baseras på prediktionsfel.

`Mallows` $C_{p'}$ `Statistika` definieras så ([7] sid 216):

$$C_{p'} = \frac{SSE_{p'}}{\hat{\sigma}^2} + 2p' - n \tag{2.14}$$

där $\hat{\sigma}^2$ är variansskattning från den fullständiga modellen med k' förklarande variabler och $SSE_{p'}$ är residualkvadratsumma från modellen med p' förklarande variabler ($p' \leq k'$).

Det kan visas ([7] sid 294) att

$$\frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} \hat{\text{mse}}(\hat{y}_i) = C_{p'}$$

där

$$\text{mse}(\hat{y}_i) = \text{Var}(\hat{y}_i) + [\text{BIAS}(\hat{y}_i)]^2$$

är medelkvadratfelet för prediktion av y_i i en bestämd modell och *bias* översätts från engelska som systematiskt fel.

Enligt Mallows har bra modeller $C_{p'} \cong p'$. I allmänhet väljs modeller med låga värde på $C_{p'}$, eftersom att minimera C'_p är ekvivalent med att välja modell som gör prediktionsfelen så små som möjligt. Men $C_{p'}$ är en stokastisk variabel och det är därför inte lätt att skilja mellan två modeller vars $C_{p'}$ värden är nästan lika.

`PRESS Statistika`.
Enligt íden av korsvalidering splittras ett datamaterial i några grupper för att använda alla grupper utom en för modellanpasning och den kvarvarande gruppen för prediktion. Det enklaste sättet att genomföra splittringen är att ta tillfälligt bort observation i ur datamängden. Denna observation ska användas för att bestämma prediktionsfel ([7] sid 217)

$$\hat{\epsilon}_{(i)} = y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_{(i)},$$

vars kvadratsumma ger en statistika som kallas *PRESS* (Predicted REsidual Sum of Squares)

$$PRESS = \sum \hat{\epsilon}_{(i)}^2 \tag{2.15}$$

index (i) markerar att observation i har utelämnats vid modellanpassningen. Bra modeller har låga *PRESS*-värden.

Ett sätt att kombinera C_p' och *PRESS* är att bestämma några kandidatregressionsmodeller mha C_p' kriteriet, vars värde för alla möjliga modeller kan beräknas på en gång genom användning av något statistiskt datorprogrampaket, och därefter räkna *PRESS* värde endast för dessa kandidatmodeller.

2.5 Inflytelsediagnostik

Observationer vars utelämning orsakar betydande förändringar i regressionsanalysen kallas för *inflytelserika* . Alla definitioner i detta avsnitt kommer från [7]. Ett mått på observationens inflytande är *Cooks avstånd,* D_i. Det finns några definitioner av D_i, här redovisas den enklaste som möjliggör att göra beräkningar för hand

$$D_i = \frac{1}{p'} r_i^2 \frac{h_{ii}}{1 - h_{ii}} \tag{2.16}$$

där p' är antal parametrar, r_i är studentiserade residualer och h_{ii} är leverage (se definitionerna nedan); Av (2.16) framgår det att stora värde på D_i kan bero på stora r_i, stora h_{ii} eller båda två.

`Leverage`, h_{ii}, ges av diagonalelementen i den så kallade *hat matrisen* H

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \tag{2.17}$$

d v s

$$h_{ii} = \mathbf{x}_i^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_i \tag{2.18}$$

där $\mathbf{X}$ är $n \times p'$ designmatrisen och $\mathbf{x}_i^t$ är radvektorn i.

Leverage återspeglar läge av $\mathbf{x}_i$ i förhållande till $\bar{\mathbf{x}}$. Undersökning av h_{ii} kan avslöja observationer som potentiellt är inflytelserika p g a deras extrema värde i $\mathbf{X}$. Observationer vars h_{ii} överskrider $2p'/n$ bör undersökas närmare.

`Studentiserade Residualer` r_i (eng. *internally studentized residuals*) återspeglar modellens anpassning i det $i$te utfallet och de fås ur de "råa" residualerna $\hat{\epsilon}_i$ genom att variansstandardisera dem genom division med deras medelfel

$$r_i = \frac{\hat{\epsilon}_i}{\sqrt{\mathrm{Var}(\hat{\epsilon}_i)}} = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{2.19}$$

Vi ser av (2.19) att stora h_{ii} leder till små $Var(\hat{\epsilon}_i)$. Alltså kommer plot över residualerna $\hat{\epsilon}_i$ inte att hjälpa att se sådana extrema observationer, ty deras residualer blir ju små. Plot över studentiserade residualer r_i är att föredra.

En fullständig analys kräver att ta hänsyn till D_i, r_i och h_{ii} för varje extremt fall.

2.6 Outlier test

Ett viktigt antagande inom regressionsanalys är att modellen är lämplig för hela datamaterialet. Observationer, som inte följer samma modell som resten av data, kallas för *'outliers'*.

Antag att $i$te observationen är en kandidat för outlier, dvs modellen för alla andra observationer är ([7] sid 114 -116):

$$y_j = \mathbf{x}_j^t \boldsymbol{\beta} + \epsilon_j \qquad\qquad j \neq i$$

men för observationen i är modellen

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \delta + \epsilon_i$$

Det framgår att kandidater för outlier är observationer med stora $|\hat{\epsilon}_i|$. Att testa om $i$te observationen är en outlier ekvivalent med att testa om $\delta = 0$. Teststatistikan kallas för `Studentiserade residualer` t_i (eng. *externally studentized residuals*). Den kallas '*externally*' eftersom observationen i inte används vid beräkningen av skattningen av σ^2:

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i \left(\frac{n-p'-1}{n-p'-r_i^2} \right)^{1/2} \tag{2.20}$$

och den är t-fördelad med n-p'-1 frihetsgrader. När observationen i testas att vara outlier genomförs det i verkligheten n test, en för varje observation. Genom att välja signifikansnivån α/n för varje test fås en total signifikansnivå som är inte mer än $n(\alpha/n) = \alpha$.

3 Statistisk analys

3.1 Beskrivning av datamaterialet

Enligt Scanias terminologi består datamaterialet av indata och utdata. Indatan utgörs av de geometriska parametrarna av intresse och utdatan är själva snurrtalet. För en detaljerad beskrivning av valet av geometriska parametrar se examensarbetet av Daniel Granquist [2].

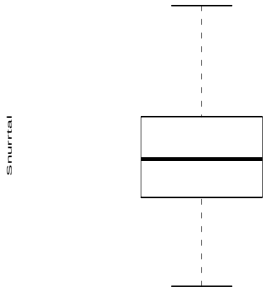
120 cylinderhuvuden valdes slumpmässigt från fyra större populationer, som har fyra olika gjutdatum. Däremot har alla 120 cylinderhuvuden bearbetats under en och samma dag. Vid början av analysen saknade två cylinderhuvud data helt och hållet, därför uteslöts de. Följaktligen finns det 118 cylinderhuvuden kvar.

3.1.1 Responsvariabeln

Responsvariabeln är *Snurrtal*, som är ett dimensionslöst tal. För att få det normerar man det varvtal som ett fiktivt paddelhjul skulle ha i cylindern med motorvarvtalet. Snurrtalet har uppmäts genom att cylinderhuvudena snurrprovades i slumpmässig ordning av Daniel Granquist. I det här datamaterialet har *Snurrtal* följande läges- och spridningsmått:

$$\min = \quad , \text{median} = \quad , \text{medel} = \quad ,$$

$$\max = \quad , \text{std.avvik.} = \quad .$$



`Figur 1.` Variationen hos *Snurrtal*

[2] D. Granquist: *Genom statistisk analys undersöka geometriska parametrars påverkan på ett XPI-cylinderhuvuds snurrtal*, KTH, VT 2010

Efteråt genomfördes åtta kontrollmätningar av snurrtal. Alltså har åtta slumpmässigt valda cylinderhuvuden två värden på *Snurrtal*. Detta kan utnyttjas för att uppskatta mätmetodens precision. Ensidig variansanalys med cylinderhuvud som kategorivariabel ger skattningen på standardavvikelsen, som är $\hat{\sigma} \approx 0.0103$.

3.1.2 Förklarande variabler

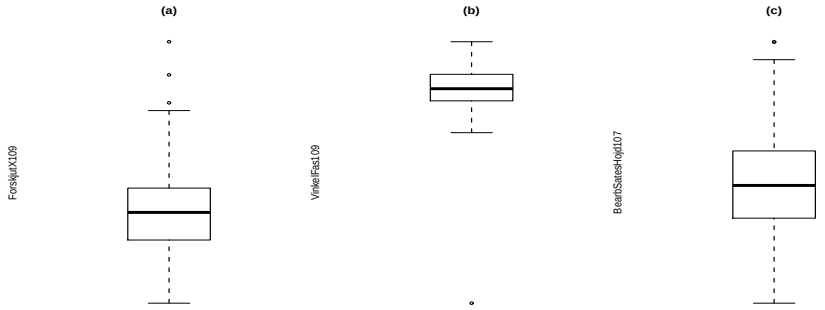
Sammanlagt finns det 35 förklarande variabler som beskriver olika geometriska parametrar på ett XPI cylinderhuvud. En stor del av dem beskriver två kanaler på ett cylinderhuvud - kanal 109 samt kanal 107. I nedanstående tabell redovisas alla variablerna:

`Tabell 1. Förklarande variablerna`

Gemensamma variabler för cylinderhuvudet	Variabler som hör till kanal 109	Variabler som hör till kanal 107
ForskjutXref	ForskjutX109	ForskjutX107
ForskjutYref	ForskjutY109	ForskjutY107
ForskjutZref	ForskjutZ109	ForskjutZ107
RakGjutkanal	RotX109	RotX107
BojdGjutkanal	RotY109	RotY107
	RotZ109	RotZ107
	ForskjutUthalX109	ForskjutUthalX107
	ForskjutUthalY109	ForskjutUthalY107
	ForskjutBearbX109	ForskjutBearbX107
	ForskjutBearbY109	ForskjutBearbY107
	HojdFas109	HojdFas107
	VinkelFas109	VinkelFas107
	BearbSatesHojd109	BearbSatesHojd107
	MinstSatesDiam109	MinstSatesDiam107
	SatesHojdD42109	SatesHojdD42107

Två av dessa variabler är kategorivariabler. De är *RakGjutkanal* (RGK) och *BojdGjutkanal* (BGK) med 3 nivåer vardera. Uppmätningen av samtliga cylinderhuvudens parametrar (förutom RGK och BGK) genomfördes av ett mätföretag med hjälp av laserskanning. Varje variabel har en och samma måttenhet, *mm*. Noggrannheten av uppmätningarna är 15 μm, dvs. avvikelsen från det sanna värdet är ± 15 μm. Ordningen på cylinderhuvudena vid uppmätningen var slumpmässig. Men i datasystemet sparas det inte de uppmätta värdena som förklarande variabler, utan det sparas skillnaden mellan det uppmätta värdet och det nominella värdet för varje variabel, dvs värdet som en variabel bör ha.

Variationen hos varje variabel har studerats både med hjälp av läges- och spridningsmått och boxplottar. Det kan sägas att alla mått har rimliga värden, vilket dock inte utesluter existens av extrema utfall (se figur 2 (b)), och spridning hos variablerna skiljer sig ganska mycket beroende på variabeln.



Figur 2. Variationen hos 3 utvalda variabler: (a)*ForskjutX109*, (b) *VinkelFas109*, (c) *BearbSatesHojd107*

Tabell 2. Läges- och spridningsmått för de plottade variablerna

	ForskjutX109	*VinkelFas109*	*BearbSatesHojd107*
min			
median			
medel			
max			
std.avvik.			

Vidare har det visat sig att det råder ett visst samband mellan två kategorivariablerna, *RakGjutkanal* och *BojdGjutkanal*, vilket kan inses mha följande kontingenstabell:

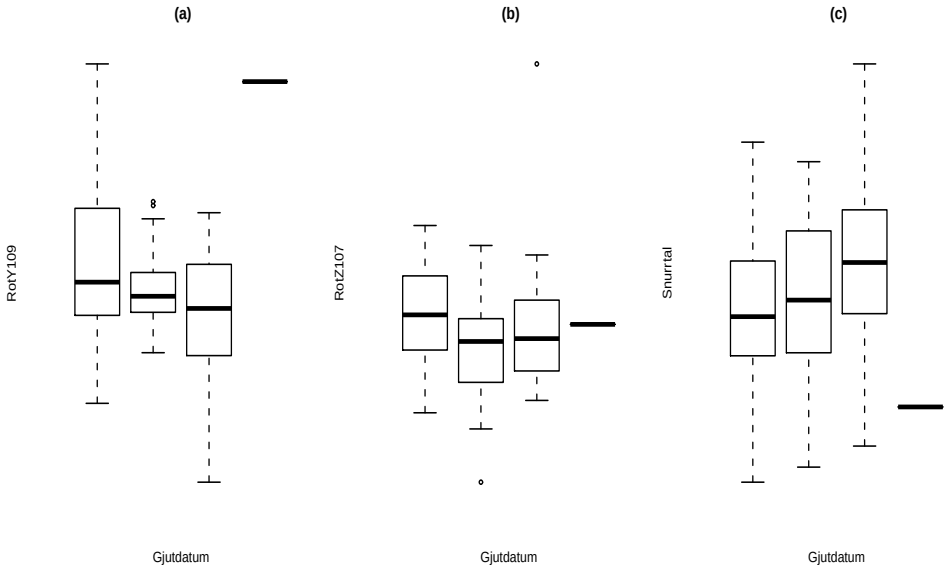
Tabell 3. Fördelningen av cylinderhuvuden över *RakGjutkanal* (RGK) och *BojdGjutkanal* (BGK).

	BGK nivå 4	BGK nivå 5	BGK nivå 6	Totalt
RGK, nivå 4	3	0	33	36
RGK, nivå 5	0	39	0	39
RGK, nivå 6	39	0	4	43
Totalt	42	39	37	118

Det framgår av tabellen att varje nivå av en variabel i de flesta fall svarar mot en bestämd nivå av den andra, dvs. informationen om den ena variabeln ger oss motsvarande information om den andra. Med andra ord, är en av dessa variabler överflödig och bör tas bort. Vi valde *BojdGjutkanal*.

3.1.3 Gjutdatum

Det var sagt tidigare att alla cylinderhuvudena var gjutna under fyra olika gjutdatum. Fördelningen har skett på följande sätt: 49 st. - dag 1, 29 st. - dag 2, 39 st. - dag 3 och 1 st. - dag 4. På Scania betraktades *Gjutdatum* ursprungligen inte som en faktor som kan ha effekt på snurrtalet. Men det är av stort intresse att undersöka om dessa fyra slumpmässiga stickprov kan slås ihop till ett stickprov. Att analysera ett stickprov i stället för några stickprov innebär att kategorivariabel av intresse inte inverkar på responsen samt på relationer mellan responsen och förklarande variabler på så sätt att olika modeller bör anpassas beroende på nivå av denna kategorivariabel. En grafisk jämförelse av variationen hos varje kontinuerlig variabel mellan olika gjutdatum har visat att några variabler har en påtaglig skillnad i spridningen mellan olika gjutdatum (se figur 3(a)). Trots detta antar vi att *Gjutdatum*s påverkan är icke-signifikant, dvs en och samma modell är giltig för alla gjutdatum. Det ska dock kontrolleras senare i analysen att detta antagande gäller för modellen, som kommer att väljas som slutgiltig.



Figur 3. Variationen hos: (a) *RotY109*, (b) *RotZ107* och (c)*Snurrtal* för fyra gjutdatum.

3.2 Fullständig modell

Efter att en kategorivariabel har tagits bort finns det 33 kontinuerliga och en kategorivariabel kvar. Dessa 34 variabler ingår alltså i den fullständiga modellen som förklarande variabler och *Snurrtal* som responsvariabel.

Samband mellan *Snurrtal* och förklarande variabler.

`Kategorivariabeln` ***RakGjutkanal*** `.`
Förekomst av kategorivariabel bland förklarande variabler i multipel linjär regression medför att analysen av sambandet mellan responsen och en kategorivariabel görs på ett annorlunda sätt än analysen av sambandet mellan responsen och en kontinuerlig förklarande variabel. Detta sätt består i *jämförelse av regressionsmodeller*. De jämförda modellerna förutsätter antingen olika modeller för varje nivå av kategorivariabel eller en och samma modell för varje nivå av kategorivariabel (se avsnitt 2.2).

Genom att beteckna alla kontinuerliga variabler med X_1, X_2, ..., X_{33} och *Snurrtal* med Y samt införa två indikator variabler Q_2 och Q_3 som betecknar nivå 5 respektive nivå 6 av kategorivariabel *RakGjutkanal* fås tre regressionsmodeller:

`Den` ***allmänna*** `modellen`

$$\begin{aligned} Y = \alpha &+ \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{33} X_{33} + \gamma_2 Q_2 + \gamma_3 Q_3 \\ &+ \rho_1 X_1 Q_2 + \rho_2 X_1 Q_3 + \rho_3 X_2 Q_2 + \rho_4 X_2 Q_3 + \ldots \\ &+ \rho_{65} X_{33} Q_2 + \rho_{66} X_{33} Q_3 + \epsilon \end{aligned}$$

`Modell` ***Parallella Regressioner***

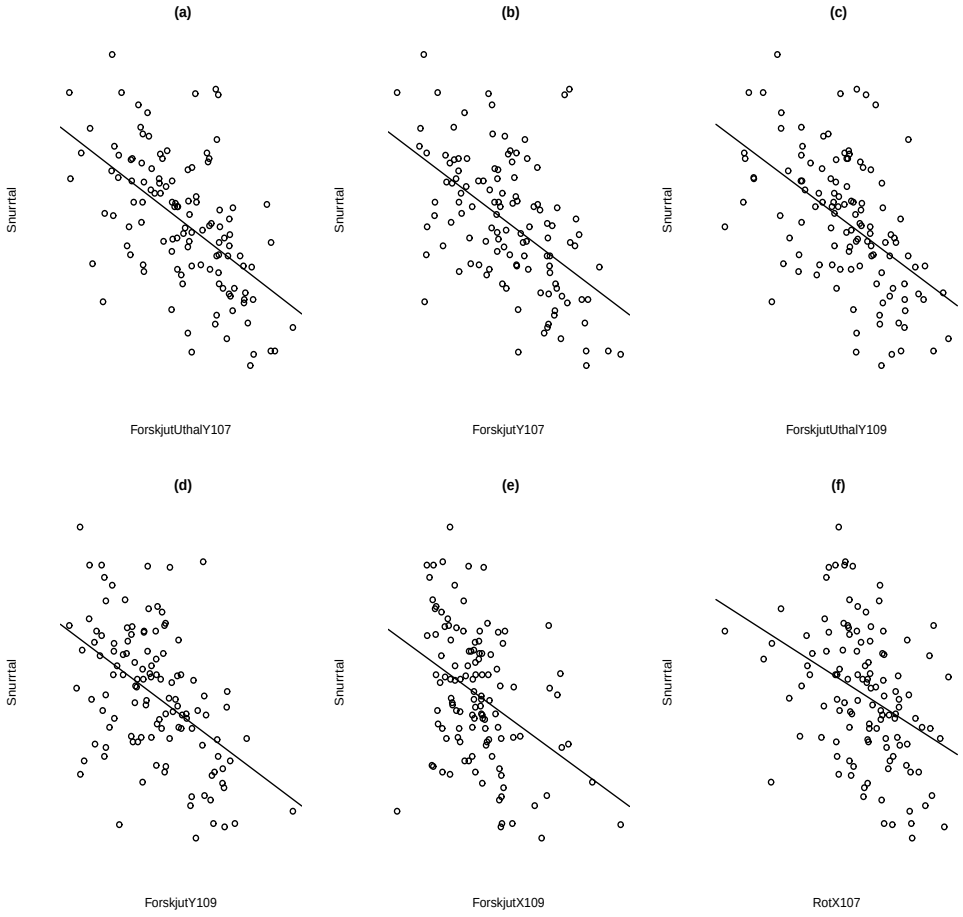
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{33} X_{33} + \gamma_2 Q_2 + \gamma_3 Q_3 + \epsilon$$

`Modell` ***Sammanfallande Regressioner***

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{33} X_{33} + \epsilon.$$

De första två modellerna förutsätter olika modeller för varje nivå av *RakGjutkanal*, medan den tredje förutsätter en och samma modell. Analysen enligt den tidigare beskrivna metoden har visat att effekten av *RakGjutkanal* är icke - signifikant. Därför har modell *Sammanfallande Regressioner* valts, som grundmodell för vidare analys.

`Kontinuerliga förklarande variabler`.
Undersökningen av hur de kontinuerliga variablerna en och en påverkar snurrtalet har visat att ganska få variabler har ett relativt starkt samband med responsen. Den variabel som har det starkaste sambandet är *ForskjutUthalY107* med korrelationskoefficienten $r = -0.557$ och $R^2 = 0.3107$ [3]. I figur 4 plottas *Snurrtal* mot de förklarande variablerna, som har det starkaste sambandet med *Snurrtal*.

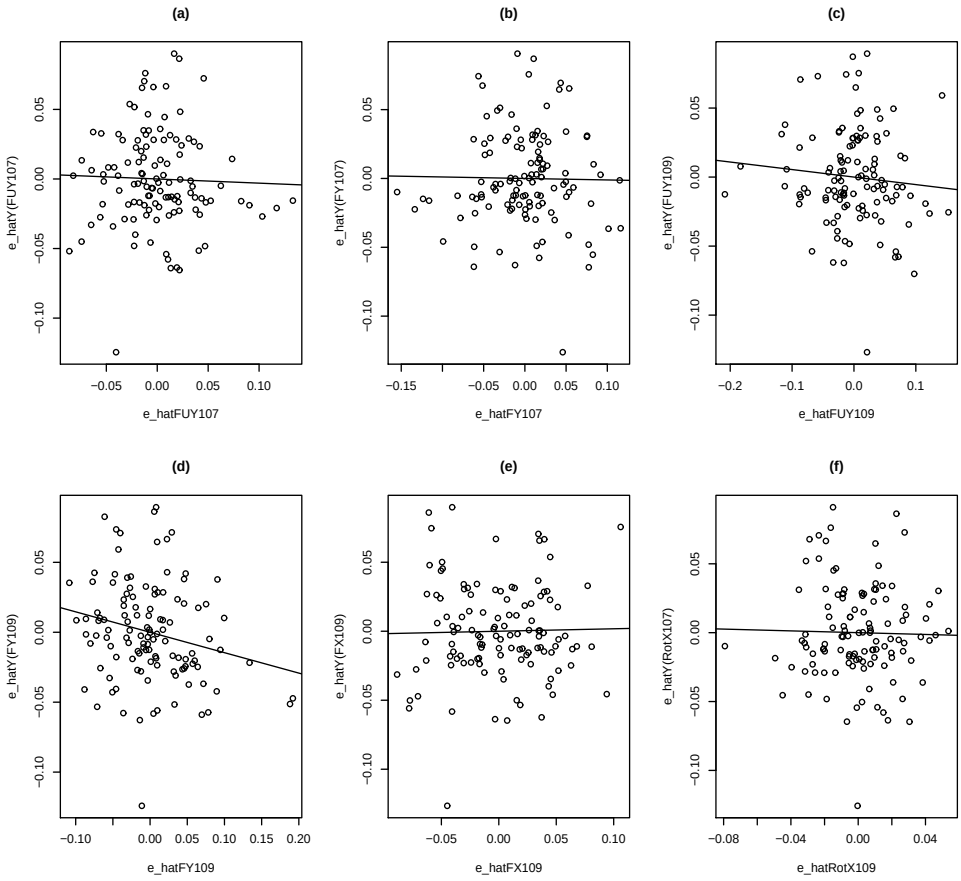


`Figur 4`. 6 scatterplottar för variabler som uppvisar ett starkt samband med *Snurrtal* med deras korrelationskoefficienter:

- (*a*) *ForskjutUthalY107*, $r = -0.56$;
- (*b*) *ForskjutY107*, $r = -0.54$;
- (*c*) *ForskjutUthalY109*, $r = -0.52$;
- (*d*) *ForskjutY109*, $r = -0.50$;
- (*e*) *ForskjutX109*, $r = -0.40$;
- (*f*) *RotX107*, $r = -0.38$.

[3] I enkel regression föreligger följande samband: $R^2 = r^2$.

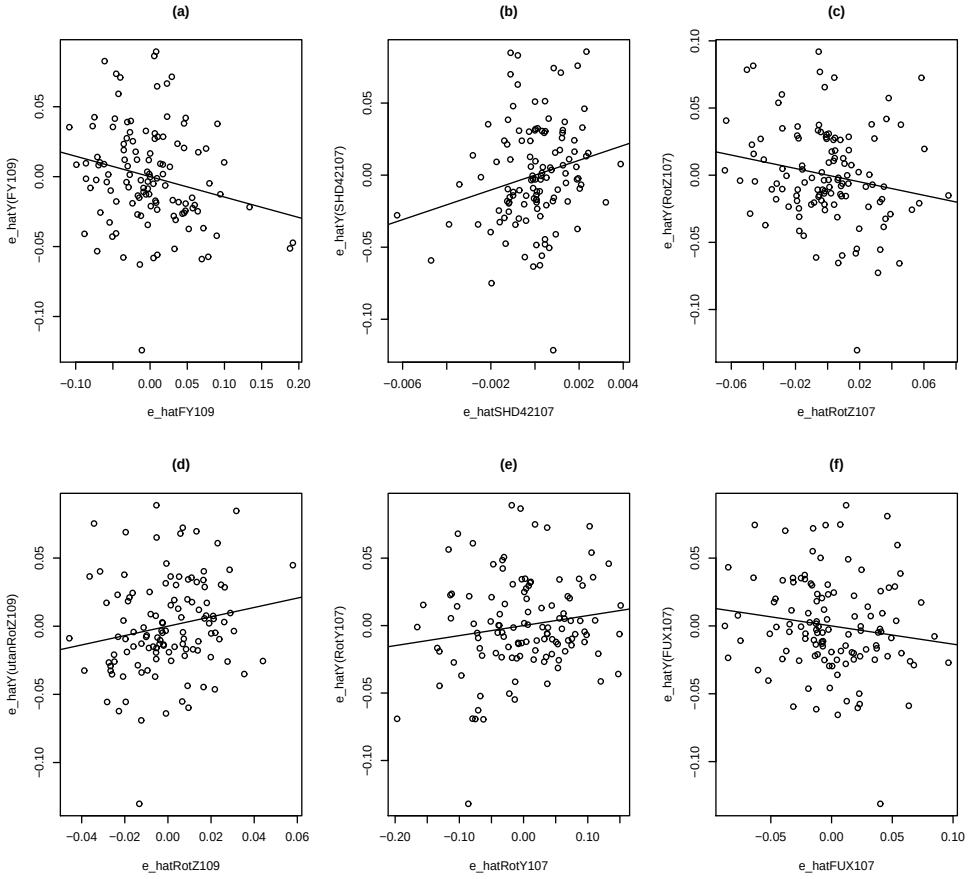
I multipel regression kan scatterplottar vara olämpliga och även missvisande pga att sambandet mellan responsen och en förklarande variabel kan påverkas av relationer mellan själva förklarande variabler och av en simultant stark inflytande på responsen av några förklarande variabler. Därför bör det studeras ett $s\ k$ justerat samband. Då tar man hänsyn till övriga förklarande variabler. Detta kan göras mha *added variabel plottar* och partiella korrelationskoefficienter. I figur 5 presenteras added variabel plottar som gäller för samma variabler vars korrelation med *Snurrtal* speglades i figur 4.



`Figur 5.` 6 added variabel plottar för variabler som uppvisat det starkaste sambandet med *Snurrtal* i Figur 4 samt deras `partiella korrelationskoefficienter`:

(a) *ForskjutUthalY107*, r^* = -0.03;
(b) *ForskjutY107*, r^* = -0.02;
(c) *ForskjutUthalY109*, r^* = -0.09 ;
(d) *ForskjutY109*, r^* = -0.22;
(e) *ForskjutX109*, r^* = 0.02;
(f) *RotX107*, r^* = -0.02.

Nu tack vare de added variabel plottarna inses det att de angivna variablernas justerade samband med *Snurrtal* är väldigt svagt jämfört med deras motsvarande icke-justerade samband. Det sagda gäller dock inte en av dessa variabler - *ForskjutY109*, vars *partiella korrelation* r^* visserligen är mindre än dess korrelation r (jämför $r^* = -0.22$ och $r = -0.50$), men den är en av de högsta partiella korrelationerna. Added variabel plottar i figur 6 gäller för variabler med de högsta *partiella korrelationerna* med *Snurrtal*, med andra ord variabler som har det starkaste justerade sambandet med responsen.



`Figur 6.` 6 added variabel plottar för variabler som uppvisar det starkaste justerade sambandet med *Snurrtal* samt deras `partiella korrelationer`:

(a) *ForskjutY109*, $r^* =$ - 0.22;
(b) *SatesHojdD42107*, $r^* =$ 0.22;
(c) *RotZ107*, $r^* =$ - 0.19;
(d) *RotZ109*, $r^* =$ 0.18;
(e) *RotY107*, $r^* =$ 0.15;
(f) *ForskjutUthalX107*, $r^* =$ - 0.14.

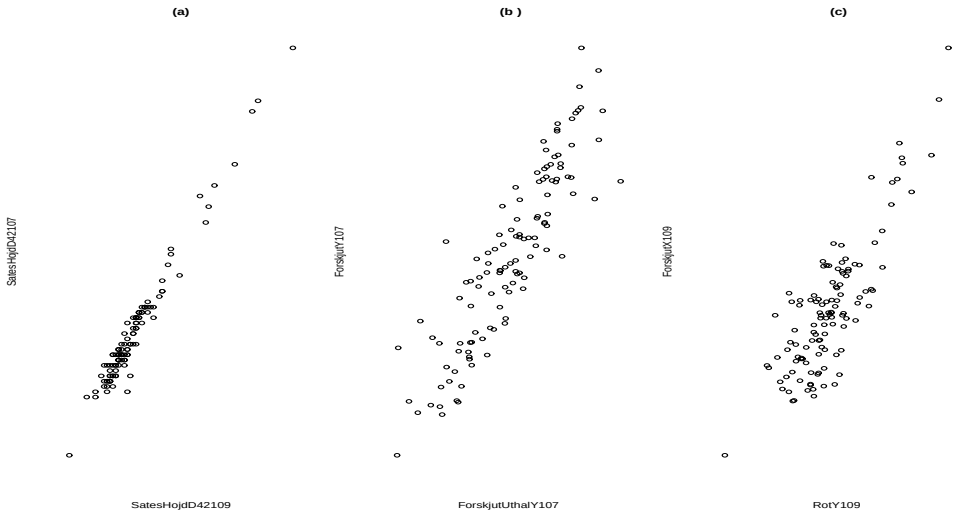
Notera att absolutbeloppet på den högsta partiella korrelationskoefficienten |0.22| inte är särskilt högt.

Förutom informationen om det justerade sambandet uppvisar added variabel plottar extrema observationer. Från alla presenterade (och icke-presenterade) added variabel plottar framgår det att en och samma observation hamnar långt bort från regressionslinjer. Detta är observation 37. Om den är 'outlier' eller någon annan extrem observation diskuteras i avsnitt Inflytelsediagnostik.

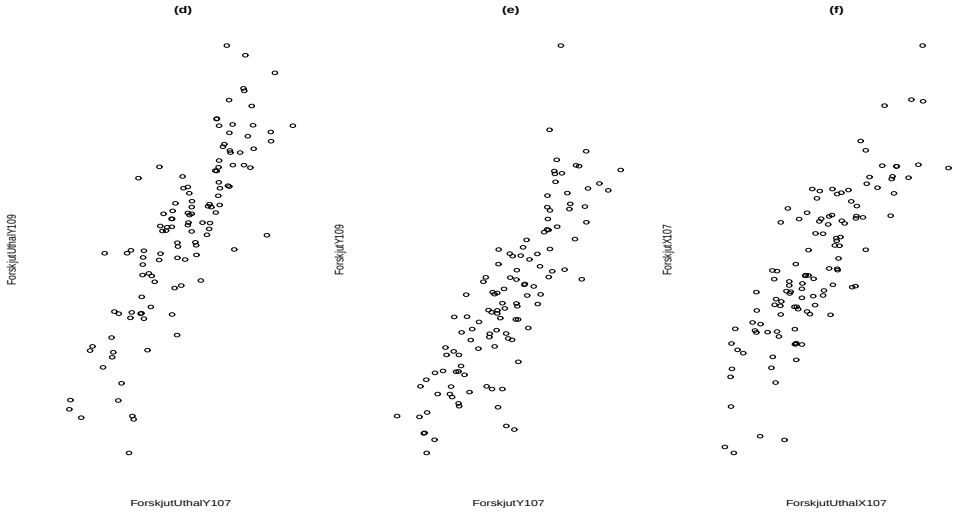
ella residualplottar med de typiska mönstren för

Korrelation mellan förklarande variabler.

Beträffande korrelation mellan de förklarande variablerna kan det sägas att kraftigt korrelerade variabler existerar. Plottarna i figur 7(a,b) gäller för variabler med hög parvis korrelation.



Figur 7a. 6 plottar, där variabler med de högsta parvisa korrelationerna plottas mot varandra:
(*a*) *SatesHojdD42109-SatesHojdD42107*, $r_{32,33} = 0.98$,
(*b*) *ForskjutY107-ForskjutUthalY107*, $r_{8,24} = 0.89$,
(*c*) *ForskjutX109-RotY109*, $r_{1,5} = 0.84$,



Figur 7b. 6 plottar, där variabler med de högsta parvisa korrelationerna plottas mot varandra:
(*d*)*ForskjutUthalY109-ForskjutUthalY107*, $r_{17,24} = 0.84$,
(*e*) *ForskjutY109-ForskjutY107*, $r_{2,8} = 0.84$,
(*f*) *ForskjutX107-ForskjutUthalX107*, $r_{7,23} = 0.84$.

Man ser att vissa variabler är nästan perfekt linjärt beroende, t ex *SatesHojdD42107* och *SatesHojdD42109* med den högsta parvisa korrelationen $r_{32,33} = 0.98$. Totalt finns det 14 par variabler vars parvisa korrelation är högre än $|0.6|$. Vad vet man om orsakerna till de observerade samvariationerna? Efter en diskussion på Scania har man kommit fram till att det kan antingen bero på tillverkningsprocessen eller på variablernas konstruktion. Som exempel betrakta två par variabler: *SatesHojdD42109- SatesHojdD42107* samt *ForskjutY107 - ForskjutUthalY107*. Angående det första paret kan man säga att sätes höjd, eller rättare sagt sätes djup, där diameter är 42 mm på båda kanalerna, bearbetades simultant på så sätt att varje förändring på den ena variabeln svarar mot nästan lika stor förändring åt samma håll på den andra. Vad gäller det andra nämnda paret kan deras starka linjära samband förklaras med variablernas konstruktion.

En sådan situation då flera variabler varierar på liknande sätt är det första tecknet på att problemet, känt som nästan multikollinearitet, kan existera. Detta innebär att överflödiga variabler kan finnas. I avsnitt 2.3 beskrivs kort de negativa effekterna som multikollinearitet medför. Tyvärr anger korrelationsmatrisen inte någonting mer än just informationen om parvis korrelation. Om flera variabler är involverade i multikolliniearitet kan dessa linjära kombinationer inte avläsas från korrelationsmatrisen. Alltså uppstår några

frågor som inte kan besvaras endast mha korrelationsmatrisen: hur grov multikollineariteten är, vilka variabler som är involverade i multikollineariteten och vilka av dem som är mest påverkade av den. Slutligen kommer vi till huvudfrågan - hur många och vilka variabler som ska tas bort för att bli av med multikollineariteten. Det bör dock sägas att det finns olika metoder att genomföra multipel regression trots existens av multikollinearitet. Bland dessa metoder är *Principal Component Regression*, *Partial Least Squares* och *Ridge Regression*. I den här analysen väljs det att eliminera vissa variabler. Information som kommer att förloras förväntas att vara minimal. I det följande avsnittet ska diagnostik av multikollineariteten för det aktuella datamaterialet och metoden för variabelelimination diskuteras.

3.3 Diagnostik av multikollinearitet och elimination av variabler

Resultat av multipel regression kan påverkas negativt av multikollinearitet. Därför är det viktigt att genomföra diagnostik av multikollinearitet innan några slutsatser angående resultat av multipel regression dras.

För att upptäcka variabler som blir påverkade av multikollinearitet undersöks VIF-värden, som för vårt datamaterial blir följande :

`Tabell 4.` VIF-värdena för den fullständiga modellen med 33 förklarande variabler.

Variabel	VIF	Variabel	VIF	Variabel	VIF
ForskjutX109	31.28	ForskjutXref	2.19	ForskjutUthalX107	7.70
ForskjutY109	28.57	ForskjutYref	1.42	ForskjutUthalY107	20.33
ForskjutZ109	11.61	ForskjutZref	2.71	ForskjutBearbX107	1.26
RotX109	8.16	ForskjutUthalX109	4.36	ForskjutBearbY107	1.44
RotY109	16.36	ForskjutUthalY109	11.96	HojdFas107	1.51
RotZ109	14.55	ForskjutBearbX109	1.33	VinkelFas107	2.28
ForskjutX107	25.06	ForskjutBearbY109	1.35	BearbSatesHojd107	1.46
ForskjutY107	22.01	HojdFas109	1.53	MinstSatesDiam109	1.39
ForskjutZ107	7.05	VinkelFas109	1.94	MinstSatesDiam107	1.51
RotX107	4.57	BearbSatesHojd109	1.43	SatesHojdD42109	56.28
RotY107	8.25			SatesHojdD42107	55.11
RotZ107	4.67				

Det framgår av tabellen att minst 10 variabler blir påverkade av multikollinearitet, om den kritiska gränsen för VIF antas vara lika med 10. Av dessa 10 variabler (till och med av alla 33 variabler) är det endast två variabler, *ForskjutY109* och *SatesHojdD42107*, som har visat sig vara signifikanta på signifikansnivån 0.05 (se tabell A1 i Appendix). Att de andra variablerna är icke-signifikanta innebär inte att de saknar betydelse för *Snurrtal* utan som sagt är det konsekvensen av multikollinearitet.

För att genomföra analys av egensystemet återvänder vi till det faktum att variationen hos de förklarande variablerna skiljer sig ganska mycket från variabel till variabel, vilket medför att de förklarande variablerna och responsen bör standardiseras enligt sambanden 2.3 och 2.4. I detta fall följer det av sambandet 2.5 att egensystemet av korrelationsmatrisen, bestående endast av korrelationer mellan de förklarande variablerna, ska analyseras.

Egenvärdena till korrelationsmatrisen blir följande

$$\lambda_1 = 4.75692 \quad \lambda_2 = 4.36503 \quad \lambda_3 = 2.97845 \quad \lambda_4 = 2.48861$$

$$\ldots\ldots$$

$$\lambda_{26} = 0.16785 \quad \lambda_{27} = 0.11891 \quad \lambda_{28} = 0.06576 \quad \lambda_{29} = 0.04976$$

$$\lambda_{30} = 0.03052 \quad \lambda_{31} = 0.01704 \quad \lambda_{32} = 0.01442 \quad \lambda_{33} = 0.00873$$

De små egenvärdena indikerar nästan linjärt beroende mellan förklarande variabler. Det finns minst sex små egenvärde i vårt fall.

Konditionstalet är

$$k = \frac{\lambda_{max}}{\lambda_{min}} = 544.638$$

vilket indikerar en ganska hög grad av multikollinearitet.

Tabell A2 i Appendix visar några egenvektorer för vår data som hör ihop med små egenvärden. Det minsta egenvärdet är $\lambda_{33} = 0.00873$, alltså är elementen i den motsvarande egenvektorn EV_{33} koefficienterna i ekvationen (2.11), där $c_0 = 0$ ty de förklarande variablerna är nu standardiserade. Genom att anta att element i denna egenvektor, som är mindre än 0.1, är approximativt lika med noll fås följande samband:

$$-0.133 \cdot ForskjutX109 + 0.691 \cdot SatesHojdD42109 - 0.682 \cdot SatesHojdD42107 \approx 0$$

vilket är ekvivalent med:

$$ForskjutX109 \approx -0.092 \cdot SatesHojdD42109 + 0.091 \cdot SatesHojdD42107.$$

På så sätt speglar elementen i EV_{33} direkt den relation som används för att generera *ForskjutX109*, *SatesHojdD42109* och *SatesHojdD42107*.

Vidare avslöjar denna egenvektor vilka variabler som bidrar mest till multikollinearitet - element som är relativt stora indikerar just sådana variabler. Det är tydligt att *SatesHojdD42109* och *SatesHojdD42107* bidrar mest. Följaktligen är dessa två variabler kandidater för elimination. För att vara säker på att förlust av informationen vid elimination är minimal tas den variabel bort som har den minsta partiella korrelationen bland alla kandidater. Alltså har vi

$$SatesHojdD42109: \quad r^* = |0.1914|$$

$$SatesHojdD42107: \quad r^* = |0.2198|$$

Variabeln *SatesHojdD42109* har ett lite svagare samband med *Snurrtal* givet att det finns 32 andra variabler i modellen. Därför tas denna variabel bort.

Efter det att *SatesHojdD42109* har tagits bort, reduceras den ursprungliga korrelationsmatrisen till en 32×32 matris . Analysen av dess egensystem (i synnerhet egenvektorn EV_{32} som hör ihop med det minsta egenvärdet) har visat att graden av multikollinearitеten har minskat något, $k = 334.136$, samt att följande variabler är potentiella kandidater för elimination: *ForskjutX109*, *RotY109*, *ForskjutY109*, *RotZ109* och *ForskjutUthalY107*. Deras nya partiella korrelationer visar vilken variabel som ska tas bort:

$$\begin{aligned} &\textit{ForskjutX109}: && r^* = |0.00176| \\ &\textit{RotY109}: && r^* = |0.05249| \\ &\textit{ForskjutY109}: && r^* = |0.20633| \\ &\textit{RotZ109}: && r^* = |0.16463| \\ &\textit{ForskjutUthalY107}: && r^* = |0.02845| \end{aligned}$$

Det blir *ForskjutX109* som elimineras. Efter eliminationen kontrolleras konditionstalet, som är $k = 275.32$, något som säger att analysen av egensystemet av den reducerade 31×31 korrelationsmatrisen är nödvändig. Proceduren av elimination upprepas på det beskrivna sättet tills konditionstalet och alla VIF-värden för alla ingående variabler blir tillräckligt små. Detta möjliggör anpassning av modellen multipel linjär regression. Resultatet av alla steg av variabeleliminationen sammanfattas i nedanstående tabell:

`Tabell 5`. Sammanfattning av variabeleliminationens procedur.

Variabel som eliminerats	Kond.talet k och VIF_{max} efter att den nämnda variabeln eliminerats	
-	544.64	56.28
SatesHojdD42109	344.14	30.85
ForskjutX109	275.32	24.47
ForskjutX107	178.21	22.37
ForskjutY107	133.57	20.15
ForskjutZ109	89.33	14.37
RotX109	49.06	6.75

Det framgår från tabell 5 att 6 variabler har eliminerats, alltså är 27 kontinuerliga förklarande variabler kvar. Med tanke på att eliminationen av variabler alltid påverkar partiella relationer mellan varje förklarande variabel och responsen, måste en ny analys av de förändrade relationerna genomföras.

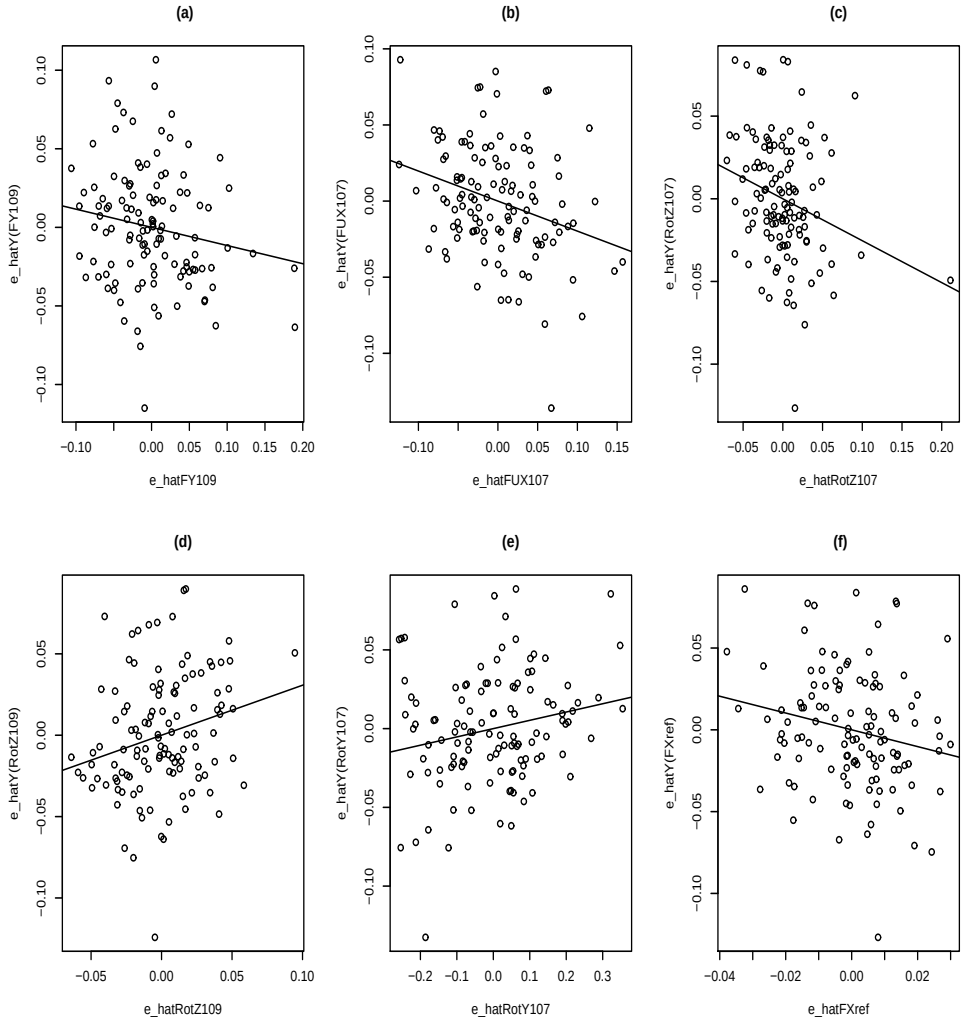
3.4 Reducerad modell

Till följd av eliminationen av 6 variabler har den reducerade modellen med 27 variabler erhållits. I detta avsnitt ska följande frågor angående denna modell tas upp:

- Sambandet mellan kategorivariabeln *RakGjutkanal* och *Snurrtal*
- Sambandet mellan varje kontinuerlig förklarande variabel och responsen justerad för övriga kontinuerliga förklarande variabler.
- Analys av behov av transformation av förklarande variabler.

`Sambandet mellan kategorivariabeln` ***`RakGjutkanal`*** `och` ***`Snurrtal.`***
Metoden för att studera det önskade sambandet består som vi vet i jämförelse av regressionsmodeller, som har beskrivits i teoridelen. Enligt analysen som i fallet med den fullständiga modellen drar vi slutsatsen att modell *Sammanfallande Regressioner*, som förutsätter en och samma modell för varje nivå av *RakGjutkanal*, är den lämpligaste.

`Justerad samband mellan` ***`Snurrtal`*** `och förklarande variabler`
Samband mellan varje kontinuerlig förklarande variabel och responsen justerad för andra förklarande variabler studeras på det sedvanliga sättet, dvs added variabel plottar och partiella korrelationer ska studeras (se efterföljande figur 8).



Figur 8. 6 added variabel plottar för variabler med det starkaste justerade sambandet med *Snurrtal* i den reducerade modellen samt deras `partiella korrelationer`:

- (*a*) *ForskjutY109*, r^* = -0.34;
- (*b*) *ForskjutUthalX107*, r^* = -0.29;
- (*c*) *RotZ107*, r^* = -0.25
- (*d*) *RotZ109*, r^* = 0.24;
- (*e*) *RotY107*, r^* = 0.20;
- (*f*) *ForskjutXref*, r^* = - 0.20

Enligt plottarna i figur 8 är det *ForskjutY109* som har det starkaste sambandet med *Snurrtal* justerad för de 26 andra variablerna. Dess partiella korrelation har blivit ännu högre efter eliminationen av 6 variabler, jämför

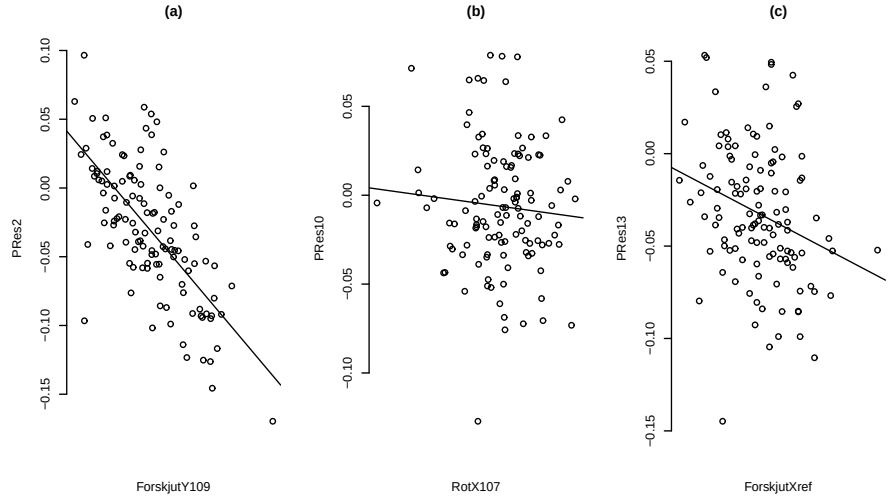
$r^*_{33\text{Var}} = -0.22$ med $r^*_{27\text{Var}} = -0.34$. Däremot har sambandet mellan *SatesHojdD42107* och *Snurrtal* försvagats från $r^*_{33\text{Var}} = 0.22$ till $r^*_{27\text{Var}} = 0.184$, vilket leder till att *SatesHojdD42107* hamnar utanför den presenterade gruppen av variablerna med de högsta partiella korrelationerna. Vad gäller andra variabler, vars added variabel plottar syns i figur 9, har deras partiella korrelationer ändrats på följande sätt:

`Tabell 6.` Exempel på förändringen av några partiella korrelationer efter att 6 variabler har eliminerats.

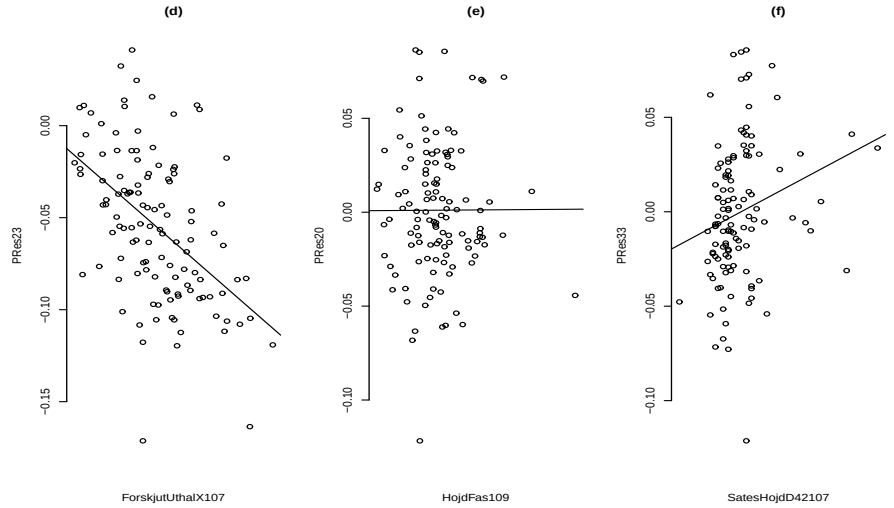
Variabel	$r^*_{33\text{Var}}$	$r^*_{27\text{Var}}$
ForskjutUthalX107	- 0.141	- 0.293
RotZ107	- 0.190	- 0.254
RotZ109	0.183	0.239
RotY107	0.155	0.204
ForskjutXref	- 0.135	- 0.202

`Transformation av förklarande variabler`

Partiella residualplottar kan visa om någon variabel behöver transformeras. Om ett sådant plott uppvisar något krökt samband, bör möjliga transformationer av motsvarande variabel analyseras närmare.



`Figur 9a.` 6 partiella residualplottar för utvalda variabler från den reducerade modellen : (*a*) *ForskjutY109*,
(*b*) *RotX107*
(*c*) *ForskjutXref*



`Figur 9b.` 6 partiella residualplottar för utvalda variabler från den reducerade modellen: (*d*) *ForskjutUthalX107*,
(*e*) *HojdFas109*,
(*f*) *SatesHojdD42107*.

Dessa plottar demonstrerar de mest förekommande mönstren i alla de 27 partiella residualplottarna. Vissa plottar uppvisar något linjärt samband, medan andra slumpmässig spridning. Angående variabeln *SatesHojdD42107* kan det antas att någon transformation kanske behövs (se figur 10(f)). En möjlig transformation är att kvadrera *SatesHojdD42107* och addera den kvadratiska termen till modellen med 27 variabler. Följaktligen fås en kvadratisk modell eller *polynommodell* med 28 variabler som är ett specialfall av den allmänna modellen för multipel regression.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{27} X_{27} + \beta_{28} X_{28} + \epsilon$$

där $X_{28} = (\text{SatesHojdD42107})^2$.

I vissa situationer är polynommodeller betydelsefulla som approximativa funktioner till okända och förmodligen väldigt komplicerade icke-linjära relationer. Men i allmänhet ska man alltid sträva efter att använda den enklaste möjliga modell som är konsistent med datamaterialet och kunskaper om problemets bakgrund.

Trots att den kvadratiska termen har visat sig vara signifikant, har jämförelsen av den reducerade modellen med de kvadratiska modellen genomförts. Resultatet av jämförelsen har visat att dessa modeller är praktiskt taget identiska

i fråga om förklaringsgrad, R^2_{adj}, graden av multikollinearitet samt uppfyllelse av modellantagandena. Jämför:

`den reducerade modellen`

$$R^2 = 0.6451 \qquad R^2_{adj} = 0.5386 \qquad k = 49.064,$$

`den kvadratiska modellen`

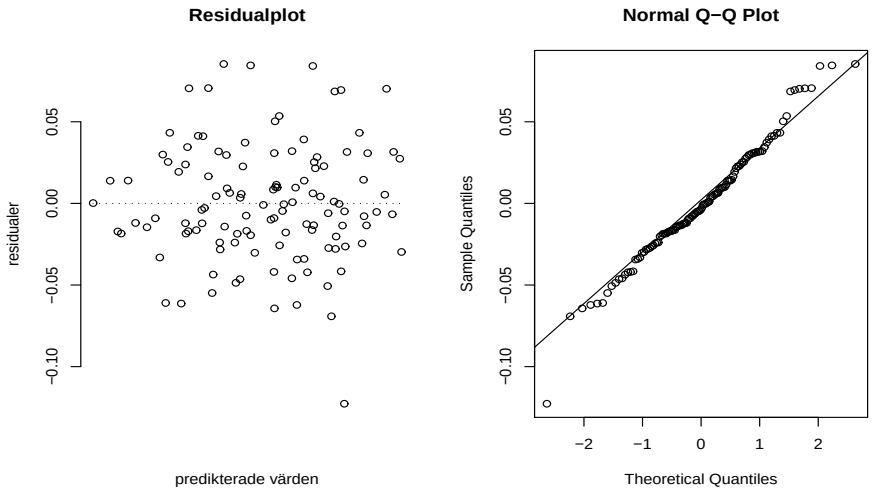
$$R^2 = 0.6609 \qquad R^2_{adj} = 0.5542 \qquad k = 53.101.$$

Dessutom avviker de båda förklaringsgraderna ytterst marginellt från förklaringsgraderna från den fullständiga modellen med 33 variabler. Jämför med:

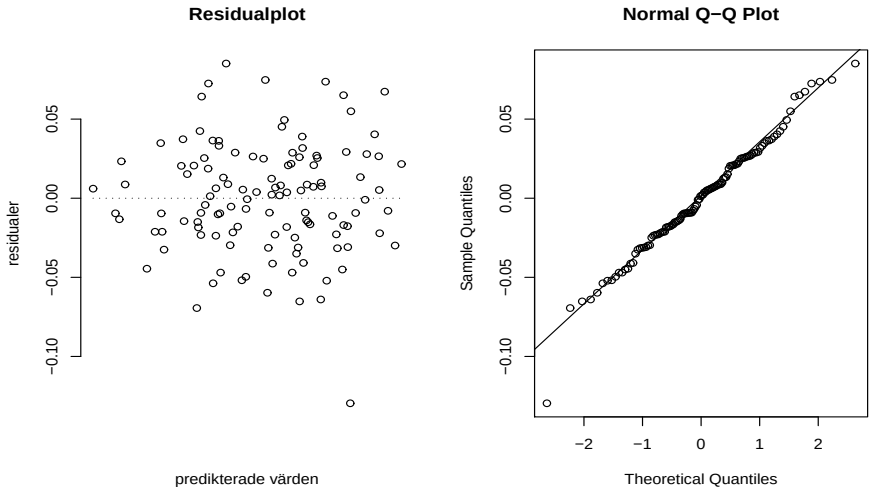
`den fullständiga modellen`

$$R^2 = 0.6602 \qquad R^2_{adj} = 0.5267$$

Enligt plottarna i figur 10 och 11 kan det sägas att modellantagandena om homogen varians och normalfördelade residualer verkar vara uppfyllda i båda modellerna.



`Figur 10.` Residualplot och normalfördelningsplot för den reducerade modellen.



Figur 11. Residualplot och normalfördelningsplot för den kvadratiska modellen.

3.5 Variabelselektion

Ett problem som uppstår nu är att välja ut ur de angivna 27 variablerna en delmängd variabler som har det största inverkan på snurrtalet. För att lösa detta problem finns det ett flertal stegvisa procedurer. De går ut på att man successivt antingen inkluderar eller eliminerar en variabel i taget tills något stoppkriterium är uppfyllt. Följande procedurer som är tillgängliga i programpaketet SAS har använts: Forward Selection, Backward Elimination, Stepwise regression (för mer information om de nämnda procedurerna se [6] sid 70-71). Resultatet av alla procedurer blir följande (risknivån för att ta in eller ta bort en variabel sattes till 5 %):

`Forward selection` och `Stepwise regression` leder till en och samma modell som innehåller följande förklarande variabler

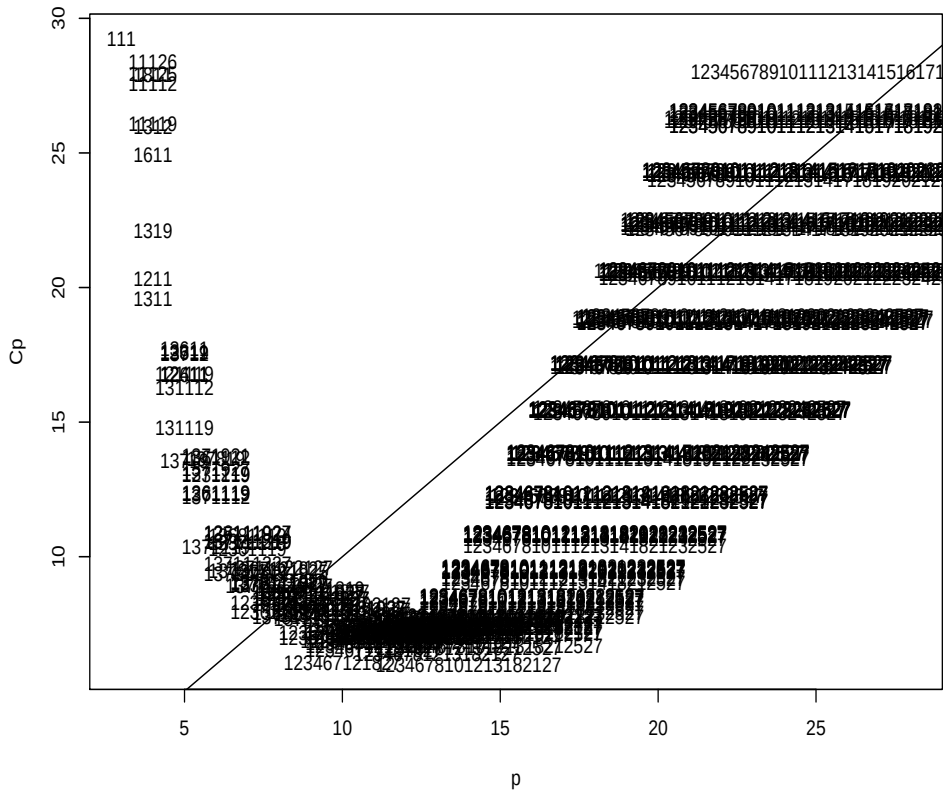
$$\textit{ForskjutUthalY107} \quad \textit{ForskjutUthalX109} \quad \textit{ForskjutY109}$$
$$\textit{RotZ109} \quad \textit{RotZ107}$$

`Backward selection` leder till modellen

$$\textit{ForskjutY109} \quad \textit{RotY109} \quad \textit{RotZ109} \quad \textit{ForskjutZ107} \quad \textit{RotY107}$$
$$\textit{RotZ107} \quad \textit{ForskjutUthalY109} \quad \textit{ForskjutUthalX107}$$
$$\textit{SatesHojdD42107}$$

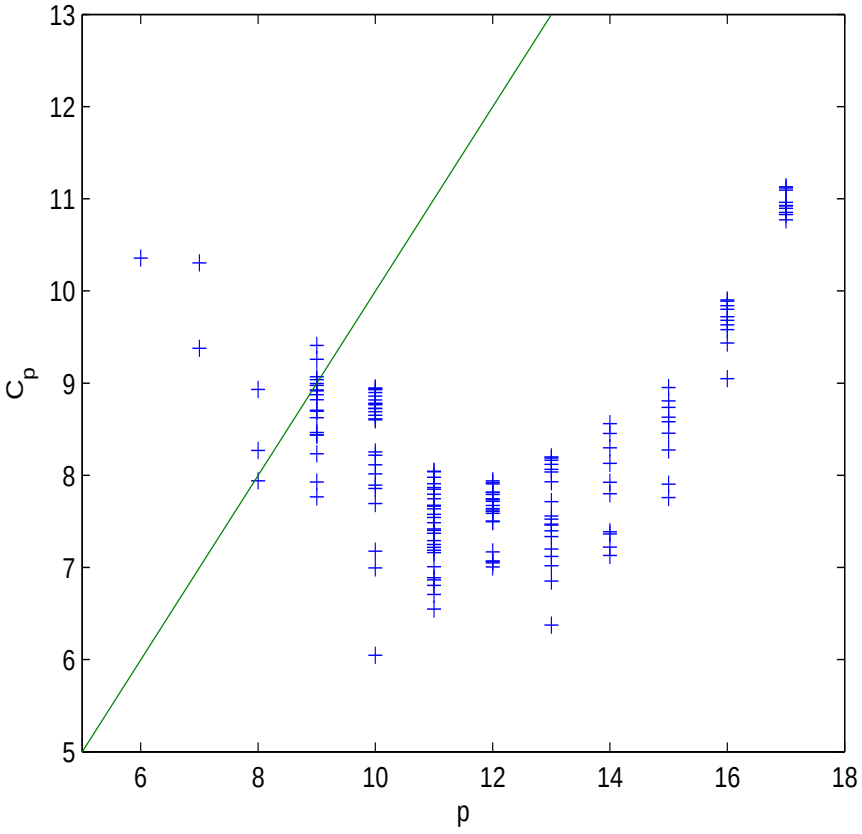
Enligt syftet av analysen ska den slutgiltiga modellen inte bara beskriva sambandet mellan responsen och de geometriska parametrarna utan den också ska ha en hög prediktionsförmåga. Därför ska andra kriterier för modellval som baseras på prediktionsfel också användas. Två sådana kriterier, $C_{p'}$ och $PRESS$, har beskrivits i avsnitt 2.4. Några kandidatregressionsmodeller bestämms först enligt $C_{p'}$ kriteriet, som kan räknas på en gång för alla möjliga modeller genom användning något statistiskt datorprogrampaket, och därefter räknas $PRESS$ värden endast för dessa kandidatmodeller.

Vid användning av $C_{p'}$ kriteriet är det hjälpsamt att konstruera $C_{p'}$ plot, där $C_{p'}$ värden plottas mot p'. Modeller med en dålig anpassning kommer att ha $C_{p'}$ som är betydligt större än p', dvs. deras $C_{p'}$ värden hamnar långt bort från linjen $C_{p'} = p'$. Bra modeller kommer att ha *låga* $C_{p'}$ värden, som ligger runt eller under linjen $C_{p'} = p'$. Notera att alla modeller på de följande $C_{p'}$ plottarna innehåller intercept, dvs $p' = p+1$ =antal förklarande variabler +1.



Figur 12. En grafisk översikt över $C_{p'}$ värden endast för modeller, med $C_{p'} < 30$. Modellerna har indikerats med index för förklarande variabler.

Enligt figur 12 hör plottens högsta $C_{p'}$ - värde till modellen som innehåller intercept, X_1 och X_{11}, där X_1 är *ForskjutY109* och X_{11} - *ForskjutUthalX109*. Det framgår också av figur 12 att det finns ett stort antal modeller med små $C_{p'}$ runt eller mindre än p'. $C_{p'}$ plotten i figur 13 visar dessa modellers $C_{p'}$ - värden tydligare.



`Figur 13.` $C_{p'}$ plot för några möjliga modeller med $6 \leq p' \leq 17$.

Enligt figur 13 finns det några modeller, vars $C_{p'}$-värden är approximativt lika med p'. Enligt Mallows, som föreslagit att bra modeller har $C_{p'} \cong p'$, kan dessa modeller betraktas som bra modeller. Å andra sidan finns det några modeller med ganska låga $C_{p'}$-värden, vilket innebär att prediktionsfelen hos dessa modeller blir små. Det lägsta $C_{p'}$-värdet hör till modellen som är resultatet av *backward elimination* ($p' = 10$). Modellen, som *forward selection* har resulterat i, har tvärtom ett högt $C_{p'}$-värde, $C_6 = 10.36$. Men detta värde är ändå mindre än några andra $C_{p'}$ värden som ligger under den räta linjen $C_{p'} = p'$, därför betraktas även denna modell som en möjlig kandidatmodell.

Vid närmare undersökning av alla möjliga kandidatmodeller har det upptäckts att de flesta modellerna innehåller icke-signifikanta variabler på signifikansnivån $\alpha = 0.05$. Men sju modeller har visat att innehålla endast signifikanta variabler. Dessa modeller med motsvarande $C_{p'}$ och *PRESS* värden

presenteras i nedanstående tabell:

`Tabell 7`. Sju möjliga slutgiltiga modeller.

p'	$C_{p'}$	PRESS				
6	10.36	0.21281	FY109 FUY107	RotZ109	RotZ107	FUX109
7	9.72	0.21378	FY109 FUY109	RotZ109 SHD42107	RotZ107	FUX109
8	7.94	0.21323	FY109 FUX109	RotY109 FUY107	RotZ109 SHD42107	RotY107
8	8.27	0.21396	FY109 FUX109	RotY109 FUY109	RotZ109 SHD42107	RotY107
9	8.99	0.21097	FY109 RotZ107	RotY109 FUY109	RotZ109 FUX107	FZ107 SHD42107
10	6.05	0.20423	FY109 RotY107 SHD42107	RotY109 RotZ107	RotZ109 FUY109	FZ107 FUX107
11	7.25	0.20454	FY109 FXref FBY107	RotZ109 FUY109 MSD109	FZ107 FBX109	RotZ107 FUX107

där F står för *Forskjut*, U - *Uthal*, B - *Bearb*, SHD42107 - *SatesHojdD42107* och MSD109 - *MinstSatesDiam109*.

I avsnittet `Gjutdatum` har det diskuterats att den slutgiltiga modellen bör kontrolleras i fråga om effekt av *Gjutdatum*, som kan inverka på så sätt att antingen olika modeller för varje nivå av *Gjutdatum* eller en och samma modell för varje nivå av *Gjutdatum* gäller. Variabeln *Gjutdatum* har därför införts i de sju möjliga slutgiltiga modellerna som kategorivariabel med 3 nivåer, dvs. dag 4, som innehåller endast en observation, har uteslutits. Därefter har metoden *jämförelse av regressionsmodeller* tillämpats på alla sju modellerna. I alla sju fallen har ett och samma resultat erhållits: effekten av *Gjutdatum* är icke-signifikant, därför har modell *Sammanfallande Regressioner*, som förutsätter en och samma modell för varje nivå av kategorivariabeln, valts. Alltså är alla modellerna kvar i analysen.

Modellantagandena om homogen varians och normalfördelade residualerna

verkar vara uppfyllda i alla modellerna, vilket har kontrollerats mha residual- och normalfördelningsplottarna (se figur A1-A7 (a,b) i Appendix)

Vad gäller *PRESS* - värdena kan det sägas att de ligger väldigt nära varandra, vilket försvårar valet av endast en modell. De större modellerna har de lägsta *PRESS* - värdena, medan de mindre modellerna är att föredra då de är enklare. Skattningen av medelprediktionsfelet, $(\text{PRESS}/n)^{1/2}$ ([7] sid 230), varierar mellan 0.04160 och 0.04258. På Scania betraktas dessa värden som tillräckligt små för att bli helt acceptabla. Detta är under antagandet att framtida observationer kommer att likna de aktuella. Att genomföra korsvalidering av modellerna på något nytt datamaterial var omöjligt pga svårigheter att genomföra nya mätningar.

Alla modeller med de standardiserade koefficienterna och anpassningsmått kan ses i tabell 8, medan det fullständiga resultatet av regressionerna kan ses i tabell A3 - A9 i Appendix. Dessutom presenteras samma modeller med icke-standardiserade koefficienter i avsnitt Inflytelsediagnostik.

Tabell 8. Sju modeller med standardiserade koefficienter och anpassningsmått

Variabel	Mod.1	Mod.2	Mod.3	Mod.4	Mod.5	Mod.6	Mod.7
ForskjutY109	-0.613	-0.664	-0.595	-0.601	-0.627	-0.588	-0.653
RotZ109	0.450	0.449	0.281	0.294	0.504	0.439	0.617
RotZ107	-0.169	-0.171	-	-	-0.232	-0.196	-0.316
ForskjutUthalX109	-0.189	-0.226	-0.231	-0.250	-	-	-
ForskjutUthalY107	-0.336	-	-0.261	-	-	-	-
RotY109	-	-	-0.168	-0.163	-0.187	-0.204	-
RotY107	-	-	0.157	0.162	-	0.192	-
SatesHojdD42107	-	0.140	0.140	0.160	0.155	0.168	-
ForskjutUthalY109	-	-0.325	-	-0.261	-0.382	-0.403	-0.376
ForskjutZ107	-	-	-	-	-0.221	-0.180	-0.218
ForskjutXref	-	-	-	-	-	-	-0.177
ForskjutBearbX109	-	-	-	-	-	-	-0.155
ForskjutUthalX107	-	-	-	-	-0.228	-0.298	-0.272
ForskjutBearbY107	-	-	-	-	-	-	-0.148
MinstSatesDiam109	-	-	-	-	-	-	0.129
R^2	0.541	0.552	0.567	0.565	0.570	0.590	0.593
R^2_{adj}	0.520	0.527	0.539	0.538	0.539	0.556	0.555

Man ser att det finns två variabler, som förekommer i alla modeller. De är *ForskjutY109* och *RotZ109*. Dessutom verkar det att det är de, som är mer viktiga för *Snurrtal* bland alla parametrar i varje modell, dvs de har större effekt på responsen i termer av standardiserade koefficienter. Emellertid måste vi komma ihåg att standardiserade koefficienter är påverkade av

variationen hos variabler, dvs om ett annat datamaterial fås med en annorlunda spridning hos variabler kan andra slutsatser dras angående hur viktiga variablerna är. Man noterar att variationen hos de standardiserade koefficienterna för *ForskjutY109* är mer stabil än variationen hos de standardiserade koefficienterna för *RotZ109*.

De näst mest förekommande variablerna är *RotZ107*, *SatesHojdD42107* och *ForskjutUthalY109* (5 modeller av 7). Bland dessa variabler verkar variabeln *ForskjutUthalY109* ha större effekt på *Snurrtal* i termer av standardiserade koefficienter. I fyra av de sju modellerna förekommer variablerna *ForskjutUthalX109* och *RotY109*.

Notera modellernas förklaringsgrader och justerade förklaringsgrader. Precis som PRESS-värdena ligger värdena relativt nära varandra inom varje mått, därmed försvåras valet av en slutgiltig modell även enligt dessa kriterier.

Konditionstalen för alla modeller varierar mellan 14.975 (modell 1) och 20.669 (modell 6), vilket indikerar att det inte finns problem med multikollinearitet i någon modell. Men det innebär inte att modellerna inte innehåller variabler med hög parvis korrelation. Här anges variabler som ingår i de utvalda modellerna och som har höga korrelationer sinsemellan:

$$\text{cor}(ForskjutY109, ForskjutUthalY109) = 0.79$$
$$\text{cor}(ForskjutY109, ForskjutUthalY107) = 0.76$$
$$\text{cor}(ForskjutY109, RotZ109) = 0.74$$

Det sistnämnda paret förekommer som sagt i alla modeller. Notera att deras korrelation inte är den högsta. Det första paret syns i fem modeller av sju och paret *ForskjutY109-ForskjutUthalY107* förekommer i endast 2 modeller.

Avslutningsvis bör det sägas att även modeller med samspelstermer har analyserats. Några modeller med högt signifikanta två-och flerfaktorsamspelstermer har erhållits. Det har dock visat sig att införandet av sådana interaktionstermer har medfört en väsentlig ökning av graden av multikollinearitet hos varje erhållna modell, vilket gör det svårt att lita på resultaten av regressionerna. Därför tas dessa modeller inte i vidare analys.

3.6 Inflytelsediagnostik

För att precisera inflytelse av någon observation med stor D_i (mått på observations inflytelse), bör en sådan observation elimineras. Resultatet av den nya regressionsanalysen kommer att visa vilka dess aspekter som har förändrats. En fullständig analys kräver även att betrakta residualer $\hat{\epsilon}_i$, studentiserade residualer r_i samt leverage h_{ii} (se avsnitt 2.5).

`Modell 1`

$$y = \alpha - 0.126 \cdot ForskjutY109 + 0.374 \cdot RotZ109 - 0.172 \cdot RotZ107 \\ - 0.099 \cdot ForskjutUthalX109 - 0.115 \cdot ForskjutUthalY107$$

`Tabell 9`. Utvalda statistikor för inflytelsediagnostik [4]

Obs. nr.	Cylin.nr	$\hat{\epsilon}_i$	h_{ii}	r_i	D_i
37	38	-0.142	0.041	-3.491	0.086
51	53	-0.062	0.184	-1.665	0.104
95	97	0.118	0.029	2.893	0.042
107	109	0.097	0.061	2.413	0.063
108	110	-0.057	0.141	-1.481	0.060
111	113	0.003	0.287	0.092	0.001

Observation 111 kunde ha en stor inflytelse pga. dess största leverage ($h_{111,111} = 0.287$). Men dess observerade inflytelse ($D_{111} = 0.001$) är liten, vilket beror på att dess studentiserade residual r_{111} är liten.

Observation 51 är inflytelserikast pga att värdet på dess leverage och r_{51} är relativt stort. Uteslutning av denna observation har påverkat mest skattningarna av *ForskjutY109*, *RotZ107* och *ForskjutUthalY107* koefficienter, fastän alla variablerna i modellen förblivit signifikanta. Förändringar i förklaringsgraden och den justerade förklaringsgraden var obetydliga i detta fall.

Observation 37 har den största residualen, som observerades även tidigare t ex på added variabel plottar; det framgår att *Snurrtal* för observation 37 är lägre än den skulle predikteras från variablerna i modell 1. Vidare är denna observation näst inflytelserikast pga dess stora värde på r_{37}, däremot är dess leverage en av de minsta, vilket pekar på att vector $\mathbf{x}_{37}$ inte skiljer sig från de andra. Anpassningen har påverkats markant genom att ta bort observation 37: $R^2 : 0.541 \rightarrow 0.586, R^2_{adj} : 0.521 \rightarrow 0.568$. Men även nu är alla variablerna siginifikanta. Outlier test (se avsnitt 2.6) har lett till att observationen 37 är outlier på den totala signifikansnivån 0.05, men ej outlier på den totala nivån 0.01.

[4] Plottar över de angivna statistikorna för alla modeller kan ses i figur A1-A7 i Appendix

Variabel *ForskjutUthalX109* blir icke-signifikant vid uteslutning av observation 107.

Modell 2

$$\begin{aligned} y = \alpha &- 0.137 \cdot ForskjutY109 + 0.415 \cdot RotZ109 - 0.175 \cdot RotZ107 \\ &- 0.119 \cdot ForskjutUthalX109 - 0.096 \cdot ForskjutUthalY109 \\ &+ 0.737 \cdot SatesHojdD42107 \end{aligned}$$

`Tabell 10`. Utvalda statistikor för inflytelsediagnostik.

Obs. nr.	Cylin.nr.	$\hat{\epsilon}_i$	h_{ii}	r_i	D_i
11	11	-0.107	0.166	-2.838	0.229
37	38	-0.145	0.043	-3.598	0.083
48	50	-0.037	0.189	-0.991	0.033
51	53	-0.078	0.137	-2.040	0.094
90	92	0.017	0.296	0.483	0.014
93	95	0.061	0.116	1.576	0.046
111	113	-0.008	0.267	-0.216	0.002

De potentiella inflytelserika observationerna 90 och 111, med deras stora leverage, har en väldigt liten observerad inflytelse. Detta beror på deras små studentiserade residualer r_{90} och r_{111}. Det är observation 11 som har den största inflytelsen p g a att både dess leverage och studentiserade residual är relativt stort. Uteslutning av detta cylinderhuvud har inte orsakat stora förändringar i regressionsanalysen med undandag av en påtaglig ökning i skattningarna av *SatesHojdD42107* och *RotZ107*s koefficienter. Detta leder dock till att variablerna blir ännu mer signifikanta. Faktor *SatesHojdD42107* blir icke-signifikant vid uteslutning av observation 51 eller observation 95, fastän förändringarna i koefficientens skattning inte är lika stora som vid uteslutning av observation 11. Anpassningen har blivit bättre efter att observation 37 har eliminerats: $R^2 : 0.552 \rightarrow 0.60, R^2_{adj} : 0.527 \rightarrow 0.577$, medan alla variablerna förblivit signifikanta. Outlier test på den totala nivån 0.05 indikerar observation 37 som en outlier, men ej på nivån 0.01.

Modell 3

$$y = \alpha - 0.122 \cdot ForskjutY109 - 0.041 \cdot RotY109 + 0.233 \cdot RotZ109 + 0.044 \cdot RotY107 - 0.121 \cdot ForskjutUthalX109 - 0.089 \cdot ForskjutUthalY107 + 0.733 \cdot SatesHojdD42107$$

`Tabell 11`. Utvalda statistikor för inflytelsediagnostik.

Obs. nr.	Cylin.nr	$\hat{\epsilon}_i$	h_{ii}	r_i	D_i
11	11	-0.066	0.263	-1.904	0.162
15	15	-0.038	0.192	-1.043	0.032
37	38	-0.144	0.077	-3.682	0.142
90	92	-0.001	0.316	-0.042	0.000
95	97	0.105	0.085	2.716	0.086
108	110	-0.068	0.189	-1.865	0.102

Uteslutning av den inflytelserikaste observationen 11 har lett till att två variabler, *RotY109* och *RotY107*, har blivit icke-signifikanta samt till en markant ökning i skattnigen av *SatesHojdD42107*s koefficient. Dess höga inflytelse beror på att både dess leverage och studentiserade residual är relativt stort. Uteslutning av observation 37, som har den största residualen, har medfört icke-signifikans av *RotY107* samt har avsevärt förbättrat anpassnigen: $R^2 : 0.564 \rightarrow 0.615$ samt $R^2_{adj} : 0.536 \rightarrow 0.591$. Elimination av observation 108 har minskat något signifikans av *RotY109* och *ForskjutUthalY107*, annars har regressionsanalysen inte blivit påverkat nämnsvärt. Uteslutning av observation 95 har lett till att *RotY107* och *SatesHojdD42107* blivit icke-signifikanta.

Outlier test har visat att observation 37 är outlier på den totala signifikansnivån 0.05, men ej outlier på nivån 0.01.

Modell 4

$$y = \alpha - 0.124 \cdot ForskjutY109 - 0.040 \cdot RotY109 + 0.244 \cdot RotZ109 + 0.046 \cdot RotY107 - 0.131 \cdot ForskjutUthalX109 - 0.078 \cdot ForskjutUthalY109 + 0.843 \cdot SatesHojdD42107$$

`Tabell 12`. Utvalda statistikor för inflytelsediagnostik.

Obs. nr.	Cylin.nr	$\hat{\epsilon}_i$	h_{ii}	r_i	D_i
11	11	-0.081	0.250	-2.295	0.220
37	38	-0.143	0.077	-3.653	0.139
48	50	-0.039	0.210	-1.079	0.039
90	92	0.002	0.327	0.051	0.0002
95	97	0.098	0.078	2.507	0.067

Observation 90 med den största leverage, dvs med den största potentialen att påverka regressionsanalysen, har nästan ingen inflytelse pga dess studentiserade residual är väldigt liten. Medan observation 11 med den näst största leverage är den inflytelserikaste observationen pga att dess både leverage och studendiserade residual är relativt stort. Uteslutning av denna observation har orsakat icke-signifikans av *RotY109* och *RotY107* på signifikansnivån $\alpha = 0.05$ samt en markant förändring i *SatesHojdD42107* koefficienten. Uteslutning av observation 37 har lett till att *RotY107* har blivit icke-signifikant samt till en förväntad förbättring av anpassningen: $R^2 : 0.565 \rightarrow 0.613$ samt $R^2_{adj} : 0.538 \rightarrow 0.588$. *RotY107* blir icke-signifikant även vid elimination av observation 95, annars påverkas regressionsanalysen inte avsevärt i detta fall.

Outlier test har visat att observation 37 är en outlier på den totala signifikansnivån 0.05, men ej outlier på nivån 0.01.

Modell 5

$$\begin{aligned} y = \alpha &- 0.129 \cdot ForskjutY109 - 0.046 \cdot RotY109 + 0.418 \cdot RotZ109 \\ &- 0.095 \cdot ForskjutZ107 - 0.237 \cdot RotZ107 \\ &- 0.113 \cdot \text{ForskjutUthalY109} - 0.133 \cdot ForskjutUthalX107 \\ &+ 0.815 \cdot SatesHojdD42107 \end{aligned}$$

`Tabell 13`. Utvalda statistikor för inflytelsediagnostik.

Obs. nr.	Cylin.nr	$\hat{\epsilon}_i$	h_{ii}	r_i	D_i
11	11	-0.081	0.225	-2.442	0.193
37	38	-0.143	0.055	-3.625	0.084
51	53	-0.076	0.143	-2.023	0.076
90	92	0.002	0.325	0.312	0.005
95	97	0.098	0.060	2.495	0.044
111	113	-0.052	0.269	0.615	0.015

Uteslutning av den inflytelserikaste observationen 11, har lett till att *RotY109* har blivit icke-signifikant samt till en markant ökning av *SatesHojdD42107*s koefficient, vilket medfört att denna variabel har blivit ännu mer signifikant. Dessutom har *RotZ109* visat sig vara känslig just för uteslutning av observation 11, fastän faktorn förblir högt signifikant. Uteslutning av den näst inflytelserikaste observationen 37 har tydligen förbättrat anpassningen utan att påverka markant signifikans hos någon av variablerna: $R^2 : 0.570 \rightarrow 0.617$; $R^2_{adj} : 0.53 \rightarrow 0.589$. Outlier test om observation 37 är en outlier visar att denna observation är outlier på den totala signifikansnivån 0.05, men ej outlier på nivån 0.01. Uteslutning av antingen observation 51 eller observation 95 har inte lett till betydande förändringar i regressionsanalysen och koefficienternas skattningar.

Modell 6

$$\begin{aligned} y = \alpha &- 0.121 \cdot \text{ForskjutY109} - 0.050 \cdot \text{RotY109} + 0.365 \cdot \text{RotZ109} \\ &- 0.077 \cdot \text{ForskjutZ107} + 0.054 \cdot \text{RotY107} \\ &- 0.200 \cdot \text{RotZ107} - 0.119 \cdot \text{ForskjutUthalY109} \\ &- 0.173 \cdot \text{ForskjutUthalX107} + 0.884 \cdot \text{SatesHojdD42107} \end{aligned}$$

`Tabell 14`. Utvalda statistikor för inflytelsediagnostik.

Obs. nr.	Cylin.nr	$\hat{\epsilon}_i$	h_{ii}	r_i	D_i
11	11	-0.072	0.255	-2.082	0.149
37	38	-0.131	0.072	-3.418	0.090
51	53	-0.080	0.145	-2.157	0.079
90	92	0.002	0.333	0.075	0.0003
95	97	0.078	0.109	2.078	0.053
111	113	0.003	0.310	0.093	0.0004
113	115	0.096	0.074	2.495	0.049

De potentiellt inflytelserika observationerna 90 och 111 har väldigt små observerade inflytelse, vilket beror på deras små studentiserade residualer. Observation 11 har däremot både relativt stor leverage och studentiserad residual, vilket leder till den största observerade inflytelsen. Utelämning av observation 11 gör att *RotY109* och *RotY107* blir icke-signifikanta samt att koefficienten för *SatesHojdD42107* ökar avsevärt, vilket dock höjer signifikansen av denna variabel. *RotY107* blir icke-signifikant även efter att observation 37 har tagits bort. Dessutom förbättras då anpassningen: $R^2 : 0.590 \rightarrow 0.63$ samt $R^2_{adj} : 0.556 \rightarrow 0.60$. Outlier testet har visat att observation 37 inte är outlier varken på den totala signifikansnivån 0.05 eller på den total asignifikansnivån 0.01. Elimination av observation 51 har inte påverkat regressionsanalysen mycket, fastän variabel *RotZ107* har blivit ännu mer signifikant.

Modell 7

$$y = \alpha - 0.134 \cdot ForskjutY109 + 0.513 \cdot RotZ109 - 0.094 \cdot ForskjutZ107 - 0.323 \cdot RotZ107 - 0.546 \cdot ForskjutXref - 0.111 \cdot ForskjutUthalY109 - 0.895 \cdot ForskjutBearbX109 - 0.159 \cdot ForskjutUthalX107 - 0.770 \cdot ForskjutBearbY107 + 0.274 \cdot MinstSatesDiam109$$

`Tabell 15`. Utvalda statistikor för inflytelsediagnostik.

Obs. nr.	Cylin.nr.	$\hat{\epsilon}_i$	h_{ii}	r_i	D_i
37	38	-0.119	0.094	-3.135	0.092
51	53	-0.088	0.168	-2.413	0.107
54	56	0.051	0.211	1.441	0.050
61	63	-0.008	0.255	-0.226	0.002
109	111	0.095	0.042	2.422	0.024
107	109	0.075	0.095	1.963	0.037
111	113	0.025	0.294	0.743	0.021

ForskjutXref blir mest påverkad vid elimination av den inflytelserikaste observationen 51, fastän alla variablerna förblir signifikanta. Elimination av observaion 37 medfört icke-signifikans av *ForskjutBearbX109* samt en förbättrad anpassning: $R^2 : 0.59 \rightarrow 0.63, R^2_{adj} : 0.55 \rightarrow 0.59$. Outlier testet har visat att observationen 37 inte är outlier varken på signifikansnivån 0.05 eller 0.01. Utelämning av observation 54 orsakar anmärkningsvärda förändringar i *ForskjutBearbX109* och *ForskjutBearbY107*s koefficienter, vilket dock inte påverkar variablernas signifikans.

Sammanfattningsvis kan det sägas att alla modellerna har visat att vara mer eller mindre ostabila och känsliga för elimination av endast en observation. Variabler, som är mest känsliga och blir ofta icke-signifikanta, är variablerna som har med rotation att göra, särskilt rotation i y led för båda kanalerna (*RotY109*, *RotY107*). Vidare kan noteras variabeln *SatesHojdD42107*, som i de flesta fall förblir signifikant men vars koefficientskattning kan ändras kraftigt pga elimination av en enda observation.

Observation 37 förekommer i alla modellerna, som en av de mest inflytelserika observationerna, och dess inflytelse speglas i första hand i en ökning av förklaringsgraden i alla modeller vid eliminationen av denna observation. Med tanke på de alla ganska låga förkaringsgraderna, som modellerna har, är det en önskvärd förbättring. I fem modeller av sju förekommer observation 11 och 51, som uppvisar stor inflytelse genom avsevärda förändringar i koefficienternas skattningar vid elimination av någon av dessa observationer. Efterföljande undersökning har dock misslyckats att avslöja någon

befogad orsak för elimination av de nämnda observationerna, sådana som indatafel. Men en djupare undersökning har visat att en möjlig orsak till den höga inflytelsen av observation 11 kan vara utloppshålets ”orundhet”, som kan påverka snurrtalet. Dessutom innehåller cylinderhuvudet 11 en av de extrema mätningarna för *SatesHojdD42107*, något som förklarar markanta förändringar i dess koefficientskattning vid elimination av detta cylinderhuvud.

4 Diskussion och slutsatser

Huvudsyftet med det här examensarbetet är att genomföra en statistisk analys för att bestämma vilka geometriska parametrar som påverkar snurrtalet på ett XPI cylinderhuvud. Resultatet av denna analys har dock visat att det är omöjligt att få fram en enda statistisk modell där geometriska parametrar med stark inverkan på snurrtalet presenteras, i alla fall inte med det aktuella datamaterialet. En av de möjliga orsakerna till den här osäkerheten vid modellvalet är att olika kriterier för modellval har används. Olika kriterier kan leda till olika slutsatser. Dessutom har alla presenterade modellerna visat nästan lika bra resultat enligt de olika kriterierna, vilket också försvårar valet av en enda modell. Vidare är det viktigt att framhäva det faktum att det aktuella datamaterialet är resultat av ett okontrollerat experiment. För att säkert kunna säga vilka faktorer som inverkar så skulle det behöva göras ett kontrollerat experiment, t ex ett faktorförsök, då varje parameter får variera endast på två nivåer. Med 33 parametrar och två nivåer på varje blir det mer än åtta miljarder försökspunkter (2^{33}). Med endast 118 okontrollerade försökspunkter ska vi inte förvänta oss att vi kan dra några säkra slutsatser.

Analysen av det aktuella datamaterialet har upptäckt ett samband mellan två kategorivariabler: *RakGjutkanal* och *BojdGjutkanal*. Att en nivå på den ena variabeln hör ihop med en bestämd nivå på den andra är inte ett förväntat resultat, som dessutom tros inte återspegla det sanna sambandet mellan dessa variabler. Därför finns det anledning till att betrakta detta datamaterial som ett icke-representativt stickprov för hela populationen.

En särskilld vikt bör läggas vid resultatet av anpassningen av den fullständiga modellen, i synnerhet vid den erhållna förklaringsgraden. Det är välkänt att den maximala förklaringsgraden uppnås vid regressionen på det maximala antalet förklarande variabler. Den fullständiga modellen, som innehåller alla 33 variablerna, förklarar bara 66 % av den totala variationen i *Snurrtal*, dvs ungefär en tredjedel av den totala variationen åtestår oförklarad ($R^2 = 0.6602$). I samband med detta jämför skattningen av *Snurrtal*- mätmetodens precision, $\hat{\sigma} = 0.0103$, och variansskattningen från den fullständiga modellen, $\hat{\sigma} = 0.04114$. De åtta kontrollsnurrprovningarna pekar ju på en hög precision vid alla 118 snurrmätningarna, därför är det logiskt att anta att orsaken till den relativt låga förklaringsgraden kan vara att någon eller några variabler som har det starka sambandet med *Snurrtal* inte mäts. Detta medför i sin tur den osäkerhet att hitta tillfreställande modeller, som har diskuteras tidigare. Samtidigt blir slutsatserna angående de två variablerna, som har antagits att ha större betydelse för *Snurrtal* än de andra, väldigt osäkra. De är nämligen *ForskjutY109* och *RotZ109*. Införandet av nya variabler, som kan ha starkare korrelation med *Snurrtal*, kan leda till annorlunda

slutsatser om variablernas betydelse.

Analyser av modeller som innehåller samspelstermer pekar på att möjliga variabler som kan ha en stark inverkan på snurrtal är samspel mellan variabler. Men hög grad av multikollinearitet hos analyserade modeller gör det svårt att lita på dessa resultat fullständigt. Detta stödjer förslaget att genomföra ett faktorförsök i stället för multipel linjär regression. Då ökar chansen att upptäcka eventuella samspel medan risken för misstolkningar av de observerade resultaten minskar.

Sist men inte minst har inflytelsediagnostiken av de sju möjliga slutgiltiga modellerna visat att varje modell är ganska ostabil och känslig för elimination av endast en observation. Varje modell bör kontrolleras vid insamlandet av ett nytt datamaterial. Modell 7 har förutom ostabiliteten visat på ett oväntat positivt samband mellan *MinstSatesDiam109* och *Snurrtal*. Sedan tidigare var det känt på Scania att *MinstSatesDiam109* har ett negativt samband med *Snurrtal*. Med stor säkerhet kan det dock påstås att det inte är multikollinearitet, som har orsakat det här oväntade resultatet i denna analys. Alltså finns det anledning till att studera relationen mellan denna geometriska parameter och responsen djupare.

5 Referenser

Referenser

[1] M.J.Crawley. *Statistics. An introduction using R*, Wiley, 2007.

[2] J.J.Faraway. *Practical Regression and Anova ising R.* `www.stat.lsa.umich.edu/~faraway/book`, 2002.

[3] R.J.Freund, R.C.Littell. *SAS System for Regression*, SAS Institute Inc., 1986.

[4] J.D.Jobson. *Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design*, Springer-Verlag, 1991.

[5] D.C.Montgomery, E.A.Peck. *Introduction to Linear Regression Analysis*, Wiley, 1982.

[6] R.Sundberg. *Kompendium i Tillämpad Matematisk Statistik*. Reviderad version från hösten 2009.

[7] S.Weisberg. *Applied Linear Regression*, Second Edition, Wiley, 1985.

6 Appendix

Tabell A1. Resultatet av multipel regression. Den fullständiga modellen med ostandardiserade koefficienter.

Coefficients	Estimate	Std.E	t value	Pr(>|t|)	
Intercept	.....				
ForskjutX109	0.0178	0.0922	0.193	0.8477	
ForskjutY109	-0.1460	0.0701	-2.084	0.0402	*
ForskjutZ109	-0.0386	0.0727	-0.531	0.5970	
RotX109	0.0225	0.1003	0.224	0.8234	
RotY109	-0.0386	0.0629	-0.613	0.5414	
RotZ109	0.3435	0.2016	1.704	0.0921	.
ForskjutX107	-0.0257	0.0840	-0.307	0.7600	
ForskjutY107	-0.0108	0.0768	-0.141	0.8884	
ForskjutZ107	-0.0489	0.0725	-0.674	0.5022	
RotX107	-0.0324	0.1684	-0.192	0.8479	
RotY107	0.0743	0.0517	1.438	0.1542	
RotZ107	-0.2494	0.1405	-1.775	0.0796	.
ForskjutXref	-0.3633	0.2907	-1.250	0.2148	
ForskjutYref	-0.0733	0.1377	-0.532	0.5959	
ForskjutZref	-0.2172	0.1970	-1.103	0.2733	
ForskjutUthallX109	-0.0256	0.0696	-0.368	0.7140	
ForskjutUthallY109	-0.0545	0.0651	-0.837	0.4050	
ForskjutBearbX109	-0.5442	0.4245	-1.282	0.2034	
ForskjutBearbY109	0.2473	0.4332	0.571	0.5696	
HojdFas109	0.0257	0.0813	-0.317	0.7524	
VinkelFas109	0.0646	0.1497	-0.432	0.6672	
BearbSatesHojd109	0.4947	0.5281	0.937	0.3516	
ForskjutUthallX107	-0.1342	0.1029	-1.305	0.1955	
ForskjutUthallY107	-0.0300	0.0983	-0.305	0.7609	
ForskjutBearbX107	-0.2118	0.3623	-0.585	0.5603	
ForskjutBearbY107	-0.4652	0.3957	-1.176	0.2431	
HojdFas109	-0.0600	0.0764	-0.786	0.4341	
VinkelFas107	0.1114	0.1784	0.624	0.5341	
BearbSatesHojd107	-0.6086	0.6098	-0.998	0.3212	
MinstSatesDiam109	0.1030	0.1597	0.645	0.5207	
MinstSatesDiam107	0.0579	0.1699	0.340	0.7344	
SatesHojdD42109	-4.4637	2.4967	-1.788	0.0774	.
SatesHojdD42107	5.1481	2.4926	2.065	0.0420	*

Residual standard deviation: 0.04114 on 84 degrees of freedom
Multiple R-squared: 0.6602, Adjusted R-squared: 0.5267
F-statistic: 4.946 on 33 and 84 DF, p-value: 1.692e-09

Tabell A2. Några egenvektorer för datan med 33 variabler

X_k	EV 28	EV 29	EV 30	EV 31	EV 32	EV 33
[1,]	0.430	-0.026	0.141	-0.066	0.602	-0.133
[2,]	-0.407	0.055	0.506	0.340	-0.377	0.078
[3,]	0.127	0.200	0.478	0.016	0.146	-0.016
[4,]	-0.488	-0.089	-0.164	0.085	-0.090	-0.009
[5,]	0.123	0.086	-0.119	0.125	-0.414	0.076
[6,]	0.010	0.269	-0.120	-0.092	0.341	-0.073
[7,]	0.005	-0.458	0.252	-0.526	-0.135	0.072
[8,]	0.052	0.451	0.037	-0.509	-0.132	-0.030
[9,]	-0.209	0.044	-0.259	-0.015	-0.150	0.038
[10,]	0.234	-0.054	0.114	-0.130	0.054	-0.031
[11,]	-0.046	0.191	-0.131	0.234	0.125	-0.057
[12,]	0.123	-0.279	0.021	-0.108	-0.068	0.018
[13,]	0.008	-0.025	0.020	-0.029	0.030	-0.043
[14,]	0.006	-0.053	-0.007	0.010	0.021	0.004
[15,]	-0.023	0.147	0.075	-0.051	-0.069	-0.012
[16,]	0.185	0.224	-0.117	0.019	-0.055	-0.012
[17,]	-0.160	-0.308	-0.438	-0.104	0.015	0.045
[18,]	-0.002	-0.023	-0.011	-0.022	-0.023	-0.018
[19,]	0.024	-0.037	-0.026	0.007	-0.002	0.016
[20,]	-0.070	0.033	0.016	0.010	-0.001	0.014
[21,]	0.049	0.056	0.041	0.018	0.016	0.030
[22,]	-0.003	0.048	-0.001	-0.012	-0.024	-0.021
[23,]	0.126	0.142	-0.229	0.163	-0.053	-0.053
[24,]	0.407	-0.344	0.088	0.429	0.214	-0.014
[25,]	0.018	-0.019	0.011	-0.012	-0.017	0.004
[26,]	-0.008	0.018	-0.009	-0.010	-0.009	-0.029
[27,]	0.082	0.066	-0.034	0.000	-0.003	0.007
[28,]	0.000	0.020	0.029	0.045	0.037	0.026
[29,]	-0.039	-0.023	0.019	0.007	0.016	0.015
[30,]	0.037	0.012	-0.019	-0.012	0.003	0.018
[31,]	-0.036	-0.040	0.021	0.008	0.013	0.007
[32,]	-0.020	0.018	-0.007	0.002	0.135	0.691
[33,]	-0.014	-0.108	0.020	0.005	-0.135	-0.682

Tabell A3. `Modell 1` med standardiserade koefficienter och deras signifikansnivåer.

```
Coefficients      Estimate Std. Error  t value    Pr(>|t|)
(Intercept)        1.886e-15  6.373e-02 2.96e-14 1.000000
ForskjutY109      -0.6126    0.1261    -4.857     3.9e-06 ***
RotZ109            0.4497    0.1147     3.921     0.000153 ***
RotZ107           -0.1686    0.0677    -2.491     0.014189 *
ForskjutUthalX109 -0.1886    0.0843    -2.237     0.027293 *
ForskjutUthalY107 -0.3358    0.1085    -3.096     0.002477 **

Residual standard deviation: 0.0414 on 112 degrees of freedom
Multiple R-squared: 0.5412,     Adjusted R-squared: 0.5207
F-statistic: 26.42 on 5 and 112 DF,  p-value: < 2.2e-16
```

Tabell A4. `Modell 2` med standardiserade koefficienter och deras signifikansnivåer.

```
Coefficients      Estimate Std. Error  t value  Pr(>|t|)
(Intercept)        1.876e-15  6.329e-02 2.96e-14  1.00000
ForskjutY109      -0.6644    0.1283    -5.180   1.00e-06 ***
RotZ109            0.4994    0.1165     4.286   3.89e-05 ***
RotZ107           -0.1714    0.0673    -2.546   0.01227 *
ForskjutUthalX109 -0.2262    0.0816    -2.774   0.00650 **
ForskjutUthalY109 -0.3252    0.1108    -2.934   0.00406 **
SatesHojdD42107    0.1396    0.0666     2.095   0.03843 *

Residual standard deviation: 0.04111 on 111 degrees of freedom
Multiple R-squared: 0.5516,     Adjusted R-squared: 0.5273
F-statistic: 22.76 on 6 and 111 DF,  p-value: < 2.2e-16



Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabell A5. `Modell 3` med standardiserade koefficienter och deras signifikansnivåer.

```
Coefficients      Estimate Std. Error  t value  Pr(>|t|)
(Intercept)        1.911e-15  6.251e-02 3.06e-14  1.00000
ForskjutY109      -0.5953   0.1289     -4.617  1.06e-05 ***
RotY109           -0.1683   0.0756     -2.228  0.02792 *
RotZ109            0.2808   0.1264      2.222  0.02835 *
RotY107            0.1572   0.0668      2.354  0.02036 *
ForskjutUthalX109 -0.2305   0.0811     -2.841  0.00536 **
ForskjutUthalY107 -0.2608   0.1035     -2.521  0.01314 *
SatesHojdD42107    0.1389   0.0665      2.090  0.03896 *

Residual standard deviation: 0.0406 on 110 degrees of freedom
Multiple R-squared: 0.5665,     Adjusted R-squared: 0.5389
F-statistic: 20.53 on 7 and 110 DF,  p-value: < 2.2e-16
```

Tabell A6. `Modell 4` med standardiserade koefficienter och deras signifikansnivåer.

```
Coefficients      Estimate Std. Error   t value  Pr(>|t|)
(Intercept)        1.904e-15  6.260e-02 3.04e-14  1.00000
ForskjutY109      -0.6009    0.1290     -4.659  8.94e-06 ***
RotY109           -0.1630    0.0758     -2.152  0.03359 *
RotZ109            0.2939    0.1271      2.313  0.02260 *
RotY107            0.1615    0.0670      2.411  0.01756 *
ForskjutUthalX109 -0.2498    0.0792     -3.157  0.00206 **
ForskjutUthalY109 -0.2611    0.1066     -2.451  0.01583 *
SatesHojdD42107    0.1597    0.0665      2.402  0.01798 *


Residual standard deviation: 0.04066 on 110 degrees of freedom
Multiple R-squared: 0.5652,     Adjusted R-squared: 0.5375
F-statistic: 20.43 on 7 and 110 DF,  p-value: < 2.2e-16



Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabell A7. `Modell 5` med standardiserade koefficienter och deras signifikansnivåer.

```
Coefficients      Estimate Std. Error  t value Pr(>|t|)
(Intercept)        1.904e-15  6.253e-02 3.04e-14 1.000000
ForskjutY109      -0.6269  0.1284      -4.883  3.60e-06 ***
RotY109           -0.1873  0.08523     -2.197  0.030172 *
RotZ109            0.5036  0.1226       4.109  7.73e-05 ***
ForskjutZ107      -0.2213  0.0822      -2.693  0.008210 **
RotZ107           -0.2319  0.0838      -2.769  0.006605 **
ForskjutUthalY109 -0.3822  0.1118      -3.419  0.000885 ***
ForskjutUthalX107 -0.2277  0.0830      -2.742  0.007130 **
SatesHojdD42107    0.1545  0.0663       2.330  0.021623 *

Residual standard deviation: 0.04061 on 109 degrees of freedom
Multiple R-squared: 0.5702,     Adjusted R-squared: 0.5387
F-statistic: 18.08 on 8 and 109 DF,  p-value: < 2.2e-16
```

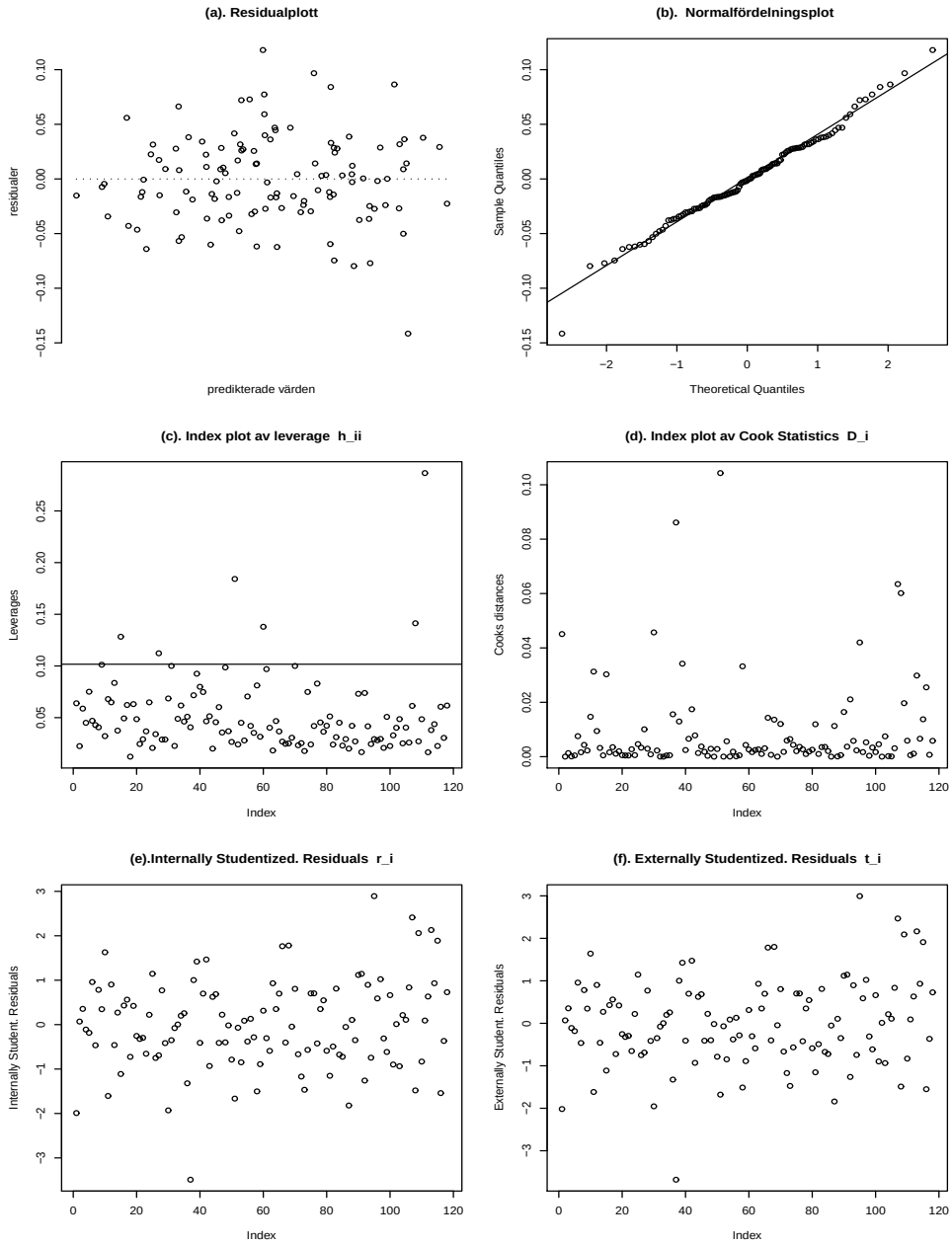
Tabell A8. `Modell 6` med standardiserade koefficienter och deras signifikansnivåer.

```
Coefficients      Estimate Std. Error  t value   Pr(>|t|)
(Intercept)        1.927e-15  6.137e-02 3.14e-14 1.000000
ForskjutY109      -0.5875   0 1272     -4.618   1.07e-05 ***
RotY109           -0.2043   0.0840     -2.432   0.016672 *
RotZ109            0.4388   0.1237      3.549   0.000575 ***
ForskjutZ107      -0.1803   0.0827     -2.180   0.031427 *
RotY107            0.1923   0.0848      2.266   0.025419 *
RotZ107           -0.1957   0.0837     -2.337   0.021306 *
ForskjutUthalY109 -0.4034   0.1101     -3.663   0.000388 ***
ForskjutUthalX107 -0.2975   0.0871     -3.415   0.000901 ***
SatesHojdD42107    0.1675   0.0653      2.565   0.011681 *

Residual standard deviation: 0.03986 on 108 degrees of freedom
Multiple R-squared: 0.5897,     Adjusted R-squared: 0.5555
F-statistic: 17.25 on 9 and 108 DF,  p-value: < 2.2e-16



Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

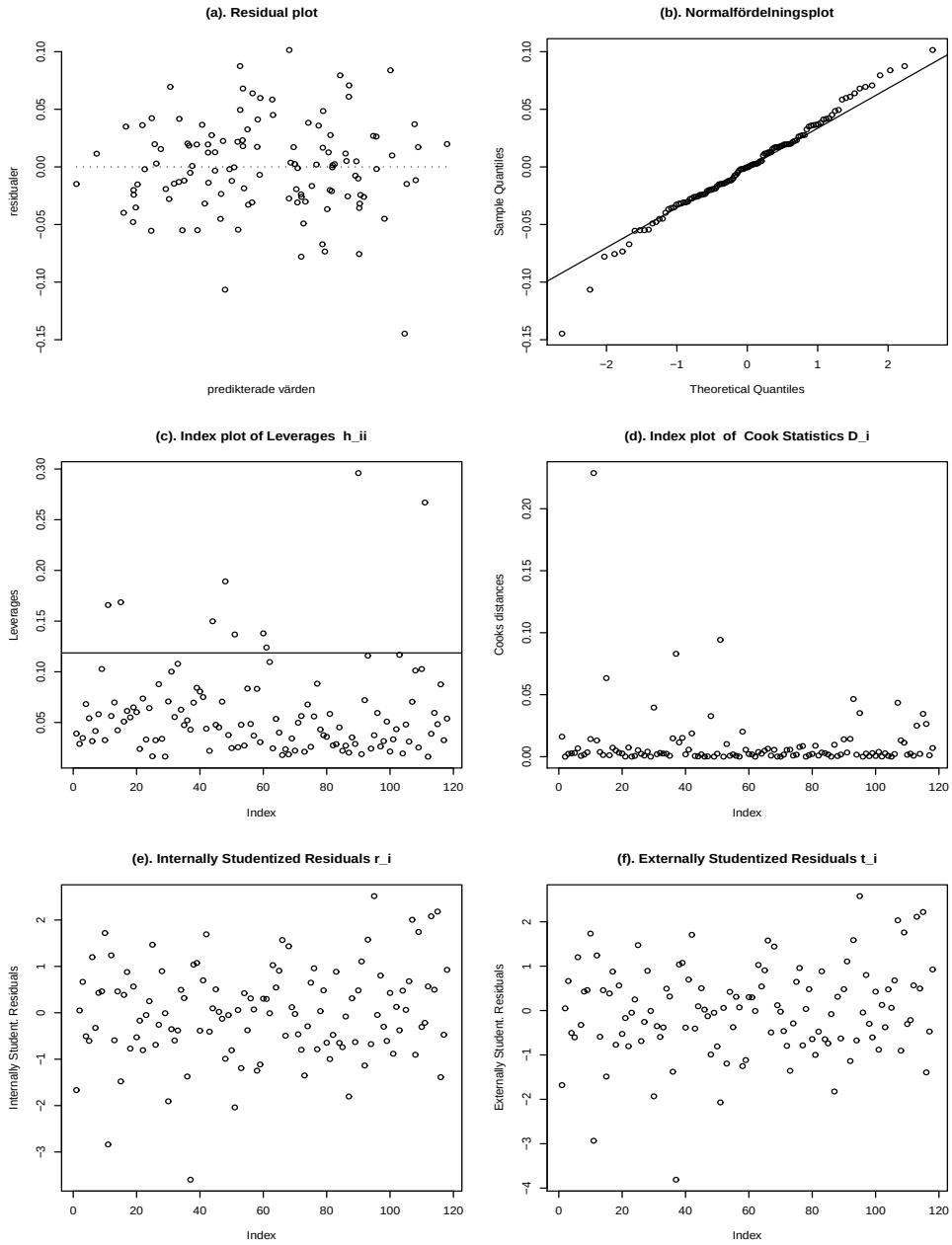
Tabell A9. `Modell 7` med standardiserade koefficienter och deras signifikansnivåer.

```
Coefficients       Estimate Std. Error  t value  Pr(>|t|)
(Intercept)       1.904e-15 6.253e-02  3.04e-14  1.000000
ForskjutY109       -0.6533   0.1234     -5.293   6.46e-07 ***
RotZ109             0.6170   0.0956      6.457   3.21e-09 ***
ForskjutZ107       -0.2179   0.0804     -2.711   0.007807 **
RotZ107            -0.3162   0.0796     -3.972   0.000129 ***
ForskjutXref       -0.1770   0.0733     -2.415   0.017450 *
ForskjutUthalY109  -0.3758   0.1103     -3.408   0.000924 ***
ForskjutBearbX109  -0.1545   0.0638     -2.421   0.017143 *
ForskjutUthalX107  -0.2723   0.0842     -3.236   0.001615 **
ForskjutBearbY107  -0.1484   0.0643     -2.307   0.022987 *
MinstSatesDiam109   0.1286   0.0632      2.036   0.044172 *

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard deviation: 0.0399 on 107 degrees of freedom
Multiple R-squared: 0.5929,     Adjusted R-squared: 0.5548
F-statistic: 15.58 on 10 and 107 DF,  p-value: < 2.2e-16
```

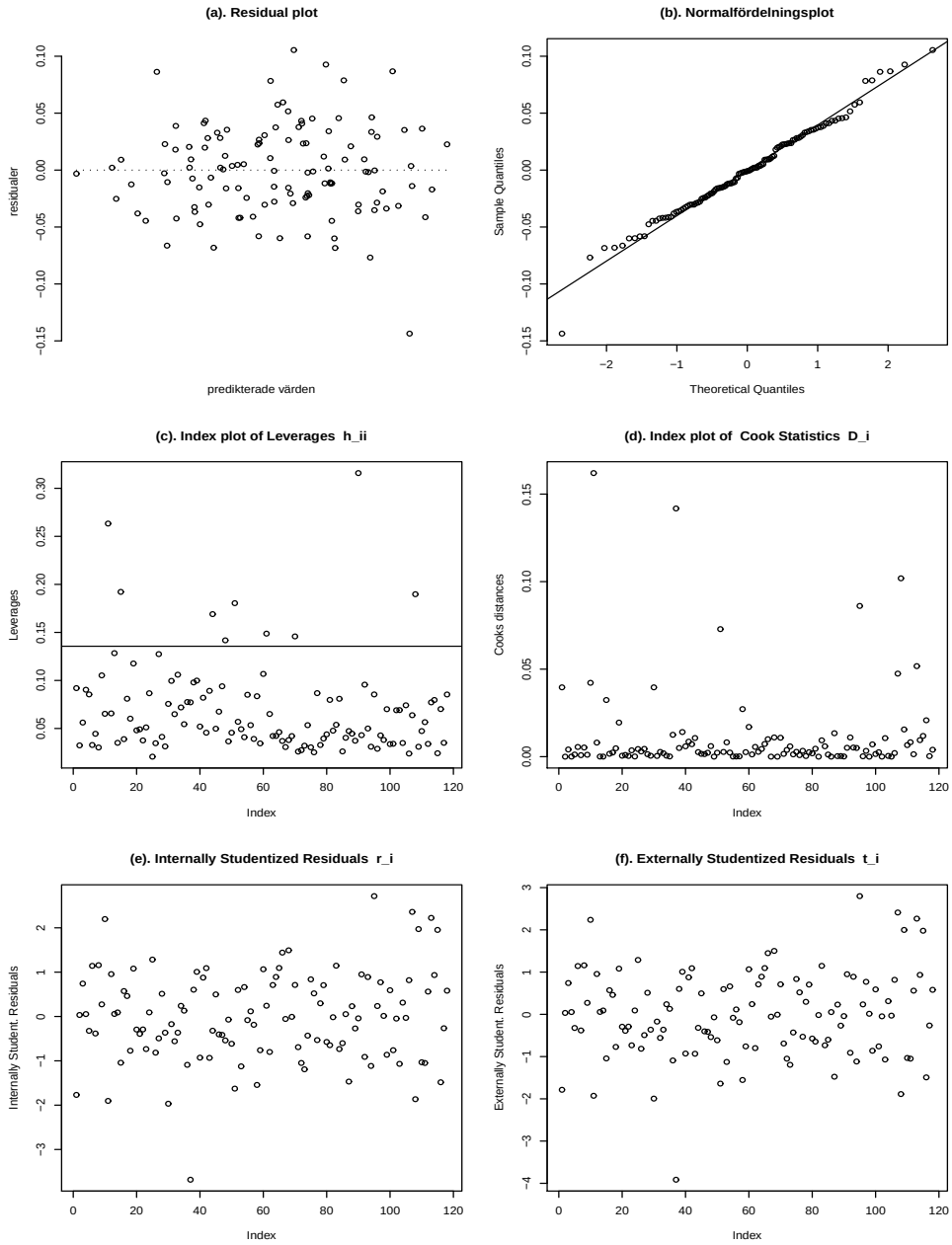
Figur A1. 6 plottar för modell 1.



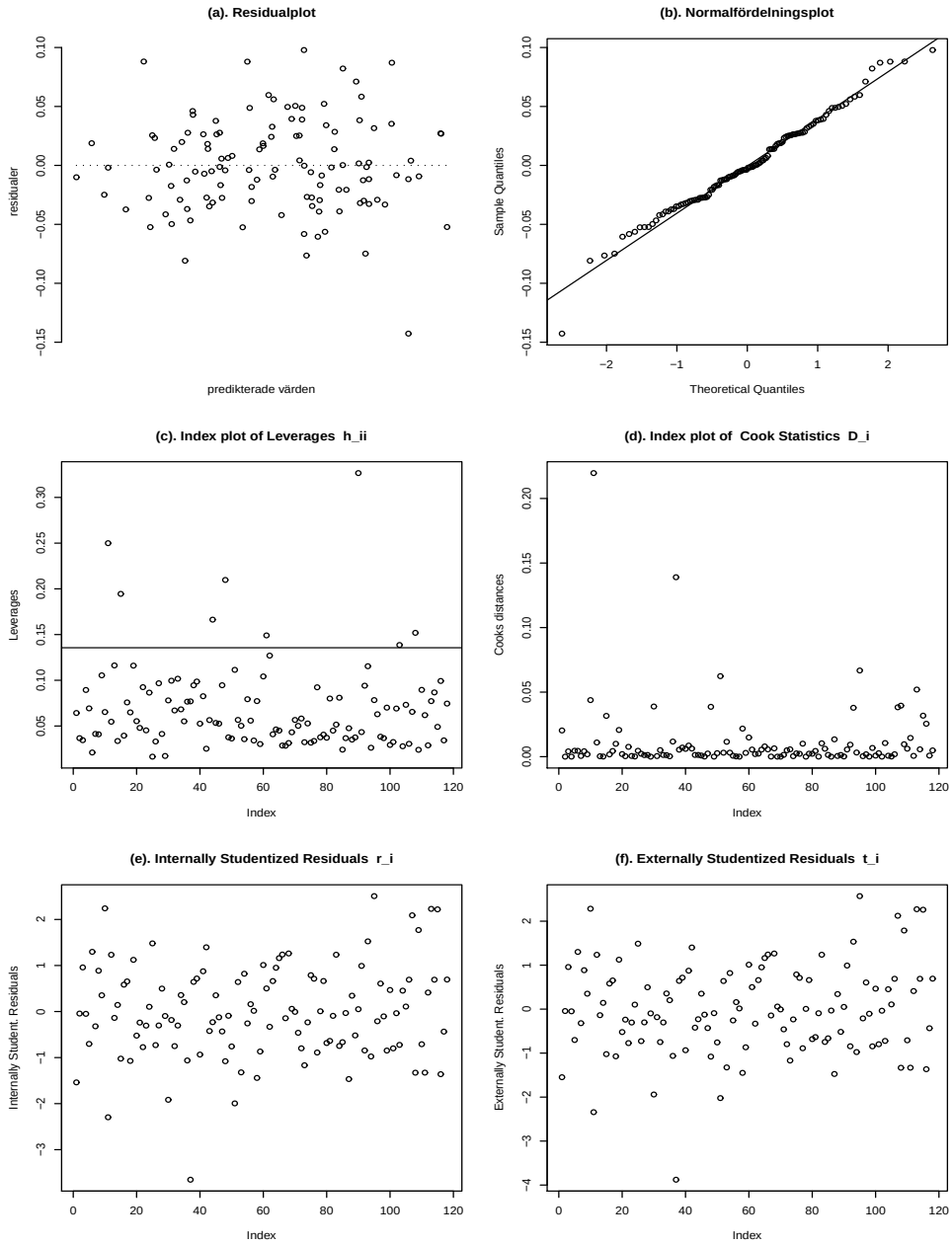
Figur A2. 6 plottar för modell 2.



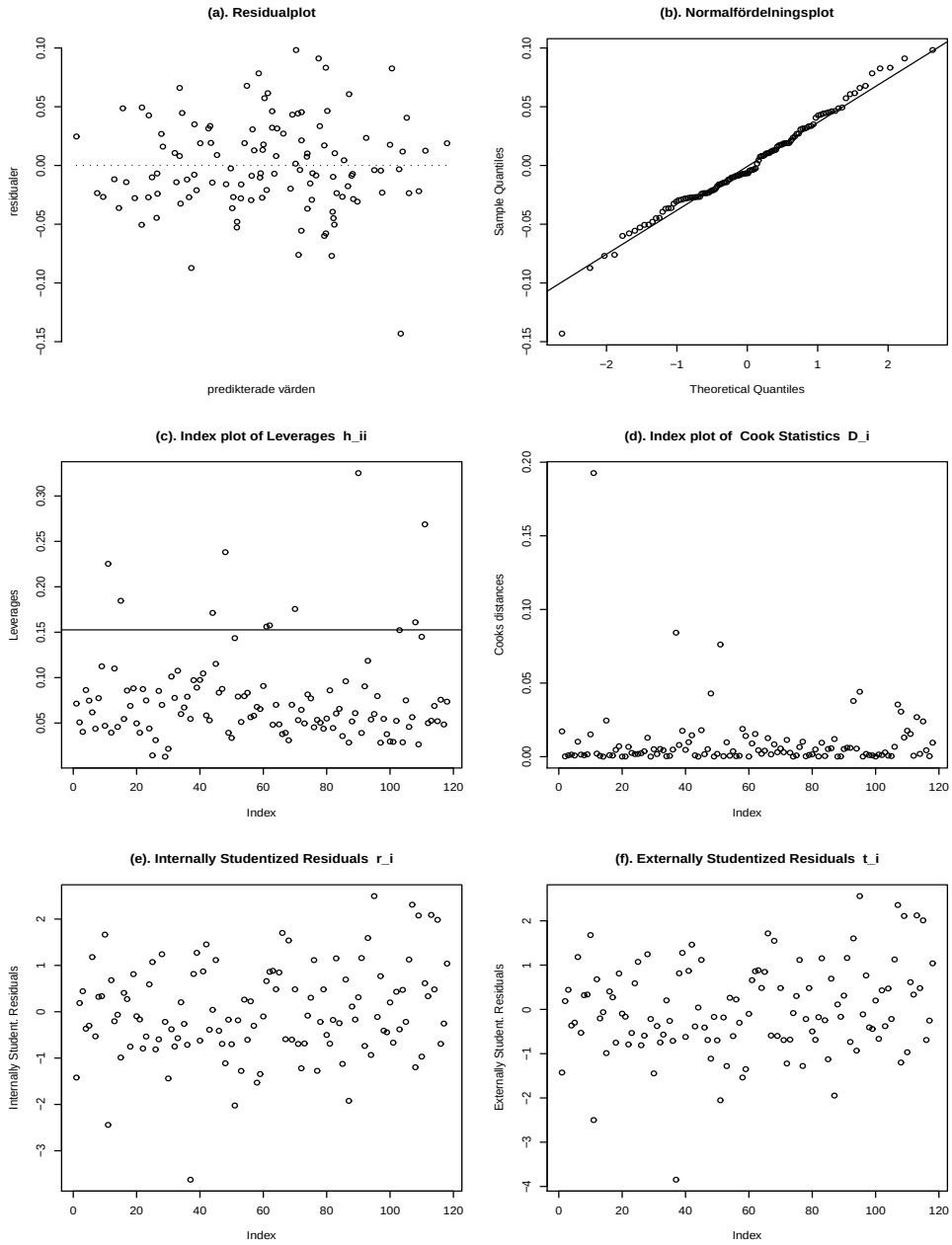
Figur A3. 6 plottar för modell 3.



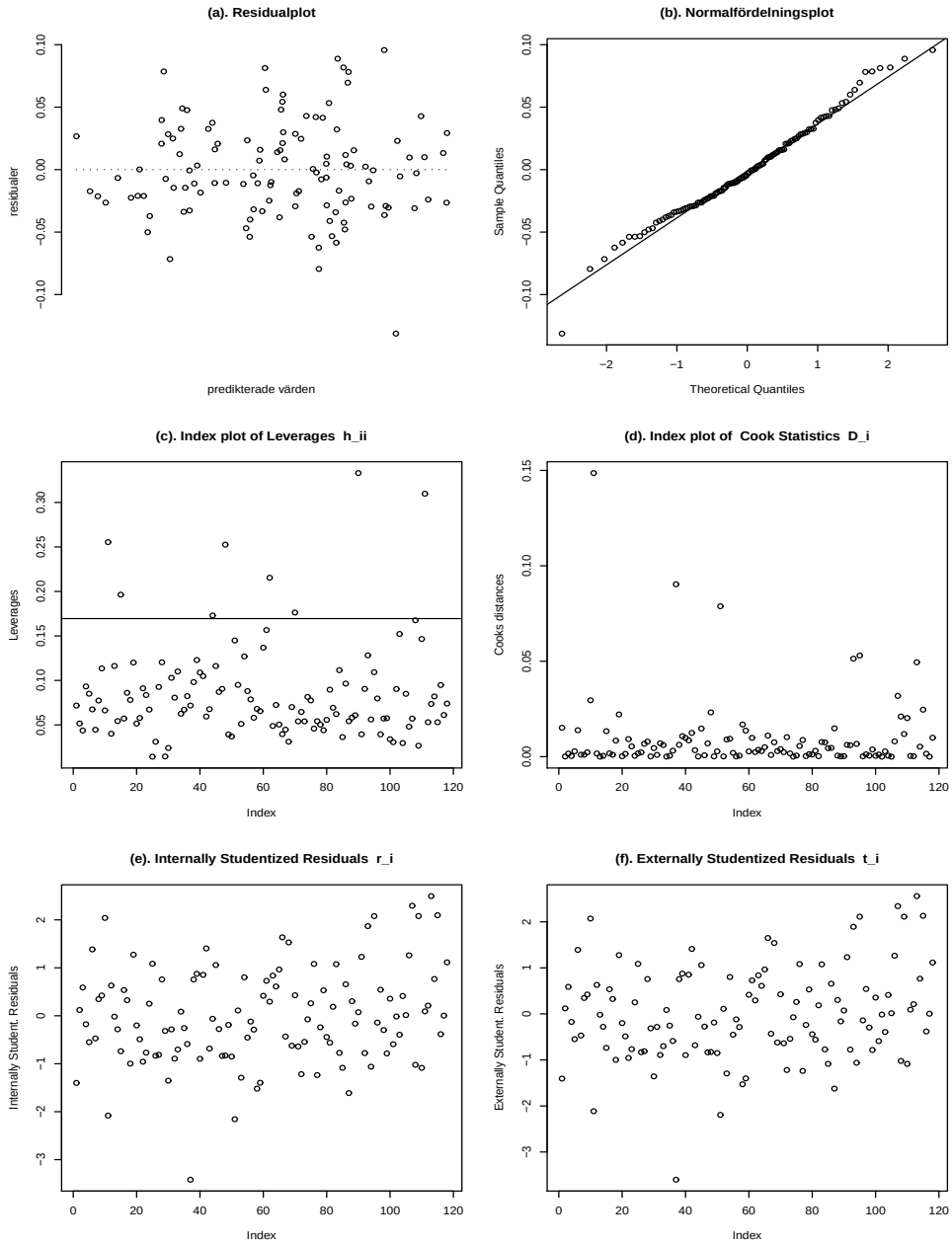
Figur A4. 6 plottar för modell 4.



Figur A5. 6 plottar för modell 5.



Figur A6. 6 plottar för modell 6.



Figur A7. 6 plottar för modell 7.