Stockholms universitet

Statistisk analys av "Friends" IMDb-betyg

Fanny Walldèn

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2024:9
Matematisk statistik
Maj 2024

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2024:9**
http://www.math.su.se/matstat

Statistisk analys av "Friends" IMDb-betyg

Fanny Walldèn*

Maj 2024

Sammanfattning

Syftet med detta arbete är att undersöka vad som påverkar hur bra ett Friends-avsnitt anses vara vilket mäts av avsnittets IMDb-betyg. Till vårt förfogande har vi ett datamaterial som bland annat beskriver allmänna egenskaper hos avsnitten såsom avsnittslängden samt huvudkaraktärernas medverkan i respektive avsnitt. Det visade sig att en multipel linjär regressionsmodell, en principalkomponent regressionmodell samt en generaliserad additiv modell kan beskriva IMDb-betyget likvärdigt. Dessa modeller visade att IMDb-betyget i allmänhet ökar genom seriens gång men minskar för vissa säsonger. Modellerna visade även att special Thanksgiving avsnitt, avsnitt som följer vissa historier samt att vissa manusförfattare och regissörer associeras med högre IMDb-betyg. Till sist ökar IMDb-betyget med antalet repliker de flesta karaktärerna har men minskar ju fler repliker Phoebe och associerade karaktärer till Phoebe har.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: fanny.wallden@telia.com. Handledare: Jan Olov Persson, Taras Bodnar.

Abstract

The goal with this thesis is to examine which factors affects how good a Friends episode is concidered to be, which is measured by its IMDb rating. To accomplish this we have a dataset that contains data such as general episode features and the involvement of the main characters in each episode. The results showed that a multiple linear regressionmodel, a principal component regressionmodel and a generalized additive model can explain the IMDb rating equivalent. These models showed that the IMDb rating generally increases throughout the serie but decreases for some seasons. The models also showed that special Thanksgiving episodes, episodes containing certain storylines and some directors and writers are associated with higher IMDb ratings. Lastly the models showed that the IMDb rating increases with the number of lines most characters has but decreases the more lines Phoebe and relatives to Phoebe has.

Förord

Inledningsvis vill jag tacka Jan Olov Persson samt Taras Bodnar. De har varit mina handledare och kommit med hjälpsamma kommentarer, råd och förklaringar på sådant jag behövt extra hjälp med. AI-verktyg har använts för att rätta kod men ej i själva uppsatsen.

Innehåll

1 Inledning

”Friends” är en komediserie som var aktiv mellan 1994 och 2004. Det har snart gått 20 år sedan sista avsnittet hade premiär men ändå har denna serie lyckats hålla sig relevant och är fortfarande, än idag, en mycket populär komediserie. Den naturliga frågan är då, hur har denna komediserie lyckats med detta? Vad är det som påverkar hur bra ett Friends-avsnitt anses vara?

Hur bra ett Friends-avsnitt anses vara kan mätas i avsnittets IMDb-betyg. IMDb står för *Internet Movie Database* och är, som förkortningen antyder, en online databas för filmer, serier och annan underhållning. IMDb användare kan lägga en röst mellan 1 och 10 på hur bra exempeltvis ett Friends-avsnitt är, dessa röster används sedan för att ge avsnittet ett IMDb-betyg mellan 1 och 10.

Vi kan slutligen formulera den primära frågaställningen som kommer besvaras i detta arbete. Vad påverkar ett Friends-avsnitts IMDb-betyg? I detta arbete kommer multipel linjär regression, principalkomponent regression, faktor analys av blandad data och en generaliserad additiv modell användas för att besvara frågaställningen.

2 Beskrivning av datamaterialet

2.1 Förberedelse av datamaterialet

Det slutgiltiga datamaterialet som kommer användas för att förklara IMDb-betyget kommer ifrån fem olika datamaterial. Nedan finns överskådlig information om dessa fem datamaterial och i appendix finns tabeller med små urval av datamaterialen.

- ifrån kaggle.com [2] kommer ett datamaterial med information om seriens alla repliker. Detta datamaterial har 67 373 rader samt 6 kolumner som visar replikerna, vem som säger dem samt när de sägs,
- ifrån github.com [3] kommer ett datamaterial med information om huvudkaraktärernas dynamiker i respektive avsnitt. Detta datamaterial har 697 rader samt 4 kolumner som visar avsnittet, dess titel samt dynamikerna i avsnittet. Variabeln som visar avsnittens dynamiker antar kombinationer av värdena $1 - 6$. Här motsvarar 1 Chandler, 2 Joey, 3 Monica, 4 Phoebe, 5 Rachel och 6 Ross,
- ifrån dedolist.com [4] kommer ett datamaterial med information om manusförfattarna och regisörerna för respektive avsnitt. Datamaterialet har 236 rader samt 8 kolumner som visar avsnittet, dess titel,

regissör, manusförfattare, premiärdatum samt antalet i Amerika som sett avsnittet,

- ifrån kaggle.com [5] kommer ett datamaterial med information om allmänna egenskaper hos respektive avsnitt. Datamaterialet har 236 rader samt 9 kolumner som visar avsnittet, dess premiärår, titel, avsnittslängd, sammanfattning, regissör, IMDb-betyg samt IMDb röster,

- ifrån thepioneerwoman.com [6] kommer information om seriens alla special Thanksgiving avsnitt. Denna information användes för att manuellt fylla i ett datamaterial med 236 rader samt 3 kolumner som visar avsnittet och huruvida det är ett special Thanksgiving avsnitt.

Datamaterialen med information om repliker, dynamik samt manusförfattare och regissörer innehåller tillsammans fyra förklarande kategoriska variabler. Respektive kategorisk variabel ändras genom att tilldela varje klass en egen variabel. Variablerna som visade avsnittens dynamiker, manusförfattare och regissörer omvandlas här till dummy variabler medan variabeln som visade vem som sa respektive replik omvandlas till diskreta variabler som visar antalet repliker respektive karaktär säger i respektive avsnitt. Resultatet av dessa ändringar är datamaterial med många förklarande variabler. Modeller med många förklarande variabler blir mer svårtolkade och kan bli överanpassade, det beslutas därmed att reducera antalet variabler. Antalet variabler reduceras genom att exkludera och i vissa fall slå samman de variabler med lite information. Slutligen slås de fem datamaterialen samman och skapar det slutgiltiga datamaterialet som kommer behandlas i detta arbete. Det slutgiltiga datamaterialet presenteras i nästa avsnitt och i appendix finns lite mer detaljer om hur de fem förberedande datamaterialen har ändrats.

2.2 Variablerna

Det slutgiltiga datamaterialet är en ihopslagning av fem datamaterial. De fem datamaterialen presenterades i föregående avsnitt och det är nu dags att presentera det slutgiltiga datamaterialet. Det slutgiltiga datamaterialet har 236 observationer där respektive observation motsvarar ett Friends-avsnitt. Det har även 26 diskreta förklarande variabler, 1 diskret responsvariabel och 2 informativa variabler. De 26 förklarande variablerna och dess innehåll är följande:

- **Avsnittsnummer:** visar avsnittets ordning under serien. Variabeln antar värden mellan 1 och 236,

- **Säsong:** visar vilken säsong som respektive avsnitt tillhör. Variabeln antar värden mellan 1 och 10,

- **Avsnitt:** visar avsnittets ordning under säsongen. Variabeln antar värden mellan 1 och 25,

- **Längd:** visar avsnittslängden i hela minuter. Variabeln antar värdena 22, 23, 24, 26, 27 och 30,

- **GaryR, KevinR, MichaelR:** är dummy variabler som visar huruvida Gary Halvorsen, Kevin Bright, Michael Lembeck eller någon annan regisserade avsnittet. Det finns således fyra klasser av regissörer: Gary, Kevin, Michael samt övriga. Gary har regisserat 54 avsnitt, Kevin 54 avsnitt, Michael 24 avsnitt och övriga 104 avsnitt,

- **AlexaM, AndrewTedM:** är dummy variabler som visar huruvida Alexa Junge, Andrew Reich och Ted Cohen eller någon annan skrev manuset till respektive avsnitt. Det finns således tre klasser av manusförfattare: Alexa, Andrew och Ted samt övriga. Alexa har skrivit manuset till 11 avsnitt, Andrew och Ted 11 avsnitt och övriga 214 avsnitt,

- **Thanksgiving:** är en dummy variabel som visar huruvida avsnittet är ett av special Thanksgiving avsnitten. Variabeln antar således två klasser: det är och det är inte ett Thanksgiving avsnitt. Det finns 10 Thanksgiving avsnitt i serien,

- **Dynamik56, Dynamik12, Dynamik13:** är dummy variabler som visar om avsnittet följer en historia om Rachel och Ross, Chandler och Joey samt Chandler och Monica. Totalt finns det fyra dynamik klasser: de tre nämnda samt att avsnittet inte följer någon av dessa historier. Det finns 70 avsnitt som följer en historia om Rachel och Ross, 36 om Chandler och Joey, 63 om Chandler och Monica samt 105 avsnitt som inte följer någon av dessa historierna. Här summeras inte klasserna till 236 eftersom vissa avsnitt följer flera av dessa historier,

- **ChandlerBing, JoeyTribbiani, MonicaGeller, PhoebeBuffay, RachelGreen, RossGeller:** visar hur många repliker respektive huvudkaraktär säger i respektive avsnitt. Chandler säger mellan 9 och 119 repliker i respektive avsnitt, Joey säger mellan 10 och 96 repliker, Monica säger mellan 4 och 77 repliker, Phoebe säger mellan 11 och 71 repliker, Rachel säger mellan 11 och 79 repliker och Ross säger mellan 7 och 87 repliker,

- **MonicaRossSläktingar, PhoebeSläktingar, RachelSläktingar:** visar antalet repliker som släktingar till dessa karaktärer säger i respektive avsnitt. **MonicaRossSläktingar** innehåller deras pappa, mostrar, kusin, farbror, Ross son Ben, Bens mamma Susan samt Ross exfru och Bens mamma Carol. Variabeln har 195 obervationer som svarar mot 0 medan resterande observationer antar värden mellan 2 och 54. **PhoebeSläktingar** innehåller Phoebes mormor, styvmamma, mamma, tvillingsyster, pappa samt svägerska. Variabeln har 218 observa-

tioner som svarar mot 0 medan resterande observationer antar värden mellan 2 och 48. **RachelSläktingar** innehåller Rachels mamma, pappa, systrar och dotter Emma. Variabeln har 224 observationer som svarar mot 0 medan resterande observationer antar värden mellan 1 och 65,

- **MikeHannigan, JudyGeller, DivPartners, Övriga:** visar hur många repliker som respektive karaktärer/ ihopslagning av karaktärer säger i respektive avsnitt. Mike är Phoebes man och har 216 observationer som svarar mot värdet 0 medan resterande observationer antar värden mellan 2 och 35. Judy är mamma till Monica och Ross och har 216 observationer som svarar mot värdet 0 medan resterande observationer antar värden mellan 1 och 23. **DivPartners** är en ihopslagning av bikaraktärer som har en relation med huvudkaraktärerna och har 128 observationer som svarar mot värdet 0 medan resterande observationer antar värden mellan 1 och 63. **Övriga** är en ihopslagning av övriga bikaraktärer och har 148 observationer som svarar mot värdet 0 medan resterande observationer antar värden mellan 1 och 33.

Responsvariabeln och dennas innehåll är följande:

- **IMDb:** visar IMDb-betyget för respektive avsnitt. Friends-avsnittens IMDb-betyg varierar mellan 7.2 och 9.7.

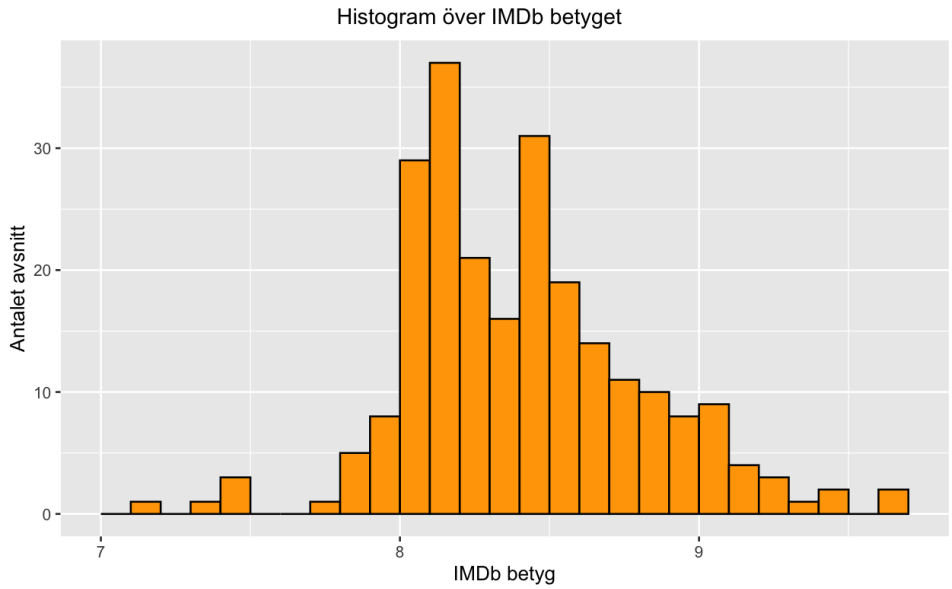
De 2 informativa variablerna och dess innehåll är följande:

- **Sammanfattning:** visar en sammanfattning av avsnittet,
- **Titel:** visar avsnittets titel.

De informativa variablerna kommer ej användas i själva modelleringen. De behålls dock för att de kan ge information som inte framgår i resterande delen av datamaterialet. Detta kan exempeltvis vara av intresse när man undersöker avvikande observationer. I avsnitt 2.3 finns histogram över IMDb-betyget. Histogram och stapeldiagram över vissa av de förklarande variablerna finns i appendix tillsammans med ett litet urval av datamaterialet.

2.3 Responsvariabeln

I det slutgiltiga datamaterialet finns det en responsvariabel, **IMDb**. IMDb är en online databas för filmer, serier och annan underhållning. IMDb användare kan lägga en röst mellan 1 och 10 på hur bra exempeltvis ett Friends-avsnitt är. Sedan används dessa röster för att, med en konfidentiell formel, beräkna IMDb-betyget.

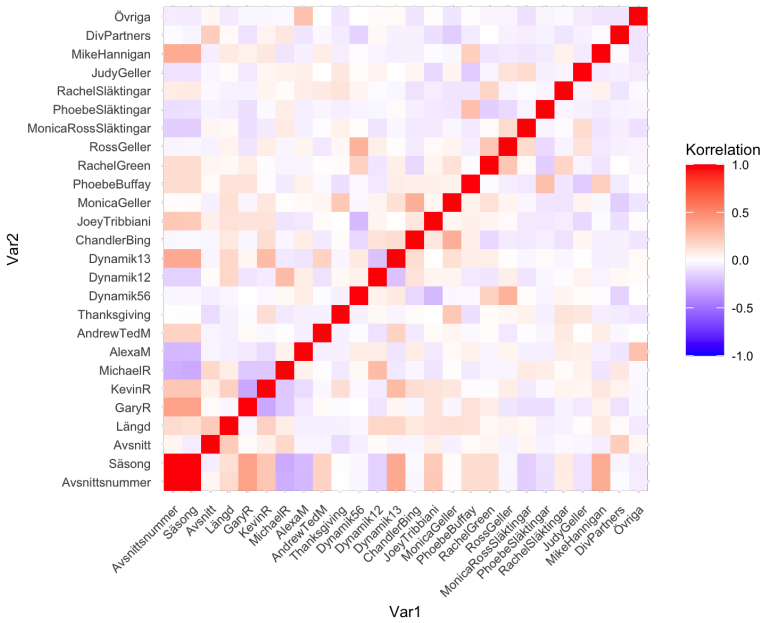


Figur 1: Histogram med klassbredd 0.1 över responsvariabeln **IMDb**

Figur 1 visar att IMDb-betyget främst ligger mellan 8 och 9. Det syns även några avvikande observationer.

2.4 Variablernas korrelation

Många förklarande variabler kan orsaka multikollinearitet. Multikollinearitet innebär att två eller fler förklarande variabler är starkt korrelerade och kan resultera i att signifikanta variabler blir insignifikanta tillsammans. Det är således viktigt att undersöka ifall några variabler är multikollineära. Vi kan direkt konstatera att avsnittets ordning inom serien har en stark korrelation med säsongen eftersom totala mängden avsnitt ökar för respektive säsong. Vi kan även konstatera att avsnittet inom säsongen, säsongen samt avsnittet inom serien är multikollineära eftersom man kan beskriva var och en av dessa variabler med hjälp av de andra två. Vi undersöker ifall det finns någon mer multikollinearitet att tänka på genom att först undersöka variablernas parvisa korrelation med Pearsons korrelationskoefficient. I figuren nedan ges en illustrering av Pearsons korrelationskoefficient.



Figur 2: Pearsons korrelationskoefficient mellan variablerna

Figur 2 visar att **Avsnittsnummer** och **Säsong** är starkt korrelerade vilket var väntat. Dessa två variabler är även enskilt korrelerade med **GaryR**, **Dynamik13** samt **MikeHannigan**. Till sist är **Dynamik56** och **RossGeller** korrelerade. Nackdelen med Pearsons korrelationskoefficient är att det bara anger korrelationen mellan två variabler. Vi vill undersöka korrelationen mellan två eller fler variabler så vi undersöker även VIF ($Variance\ Inflation\ Factor$).

Man kan läsa mer om VIF i avsnitt 3.2.5 men tills vidare nöjer vi oss med att konstatera att VIF värde > 5 indikerar multikollinearitet. Det visar sig att **Säsong** har VIF värde 66624, **Avsnitt** har VIF värde 685 och **Avsnittsnummer** har VIF värde 66507, vilket var väntat, medan resterande variablerna har VIF värden inom intervallet $[1.115, 1.625]$. Det finns således multikollinearitet mellan avsnittet inom säsongen, säsongen samt avsnittet inom serien.

2.5 Begränsningar

Datamaterialet har några begränsningar som kommer undersökas och eventuellt behandlas i kommande avsnitt. Den första och mest uppenbara begränsningen är att datamaterialet har många förklarande variabler vilket kan resultera i en svårtolkad modell som skattar parametrar dåligt samt predik-

terar ny data dåligt. Det andra som kan vara problematiskt är att det finns multikollineära variabler. Det nämndes i föregående avsnitt att **Avsnittsnummer**, **Säsong** och **Avsnitt** är multikollineära. **Avsnittsnummer** och **Säsong** är även parvist starkt korrelerade. För det tredje finns det en del dummy variabler vilka kan vara olämpliga att använda i vissa modeller, exempeltvis principalkomponent regression. Till sist finns det även ett visst bortfall i datamaterialet med replikerna. Datamaterialet med repliker har inkluderats för att få en uppfattning över hur mycket respektive karaktär är med i respektive avsnitt. Detta fungerar till viss grad men det finns karaktärer som är med betydligt mycket mer än det framstår som om man bara undersöker replikerna. Dessa karaktärer är främst barnen Ben Geller och Emma Geller Green som under stora delar av serien är för små för att kunna prata.

3 Teori

3.1 Datamaterialet

I regressionsmodeller är det önskvärt att ha så mycket data som möjligt för att hitta en modell som förklarar responsvariabeln bra. Problem kan dock uppstå ifall man har för många förklarande variabler i relation till antalet observationer i modellen. Sådana modeller ger osäkrare skattningar av parametrarna, blir mer svårtolkade och förklarar känd data bra men ny data dåligt. Utöver nämnda problem kan många förklarande variabler orsaka multikollinearitet vilket innebär att det finns minst en linjär kombination som är nästan konstant. Resultatet av multikollinearitet är bland annat att vissa signifikanta variabler blir insignifikanta tillsammans med andra. Det gäller därmed att hitta en balans mellan att bevara data och att inte ha överflödig data. I förberedandet av datamaterialet har vi haft Einstains formulering, som finns på sida 91 i *Linjära statistiska modeller* [1], i åtanke:

```
Förenkla så mycket som möjligt men inte mer än så.
```

3.2 Multipel linjär regression

Följande avsnitt om multipel linjär regression är baserad på *Linjära Statistiska Modeller* [1]. Mer specifikt är anpassningsmått baserad på sidorna 92-93, Cooks avstånd är baserad på sida 109, modellval är baserad på sidorna 95-97 och multikollineära variabler är baserad på sidorna 97-99.

3.2.1 Modellen

Modellen för multipel linjär regression skrivs:

$$Y_i = \alpha + \beta_1 \cdot x_{1i} + ... + \beta_p \cdot x_{pi} + \epsilon_i \tag{1}$$

där Y_i är responsvariabeln, $\alpha, \beta_1, ..., \beta_p$ är parametrar, $x_{1i}, ..., x_{pi}$ är de förklarande variablerna och ϵ_i är feltermen. Feltermerna ska vara oberoende och normalfördelade med konstant varians.

3.2.2 Anpassningsmått

Ett sätt att utvärdera modellen på är genom att undersöka dess anpassningsmått. Det vanligaste anpassningsmåttet är förklaringsgraden R^2. Förklaringsgraden visar hur väl den anpassade modellen beskriver datamaterialet och kan uttryckas:

$$R^2 = \frac{Kvs(regression}{Kvs(total}) = 1 - \frac{Kvs(Residual)}{Kvs(total)}. \tag{2}$$

En nackdel med R^2 är att det ökar när man tillför en variabel till modellen oavsett om den nya variabeln tillför något till modellen eller ej. Det modifierade anpassningsmåttet R^2_{adj} tar hänsyn till variablerna genom att det minskar om en variabel förbättrar modellen mindre än väntat och ökar annars. R^2_{adj} kan uttryckas:

$$R^2_{adj} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}^2_0} \tag{3}$$

där $\widehat{\sigma}^2_0 = \frac{Kvs(totalt)}{n-1}$ är σ^2 skattningen av variansen för feltermerna när man inte har någon förklarande variabel i modellen och $\widehat{\sigma}^2 = \frac{Kvs(Residual)}{n-p}$.

3.2.3 Cooks avstånd

Det finns en risk att datamaterialet innehåller avvikande observationer. Avvikande observationer kan försämra en modell vilket gör det viktigt att undersöka sådana observationer. Avvikande observationer kan identifieras med Cooks avstånd vilket mäter observationernas inflytande på modellen. Cooks avstånd skrivs

$$D_i = \frac{(\widehat{\beta}_i - \widehat{\beta})^T S(\widehat{\beta}_i - \widehat{\beta})}{m\widehat{\sigma}^2} \tag{4}$$

där $\widehat{\beta}_i$ är skattningen när observation i exkluderats, m är antalet förklarande variabler, σ^2 är standardavvikelsen för feltermerna och S är avbildningsmatrisen. Det finns flera olika gränser man kan använda för att avgöra vilka observationer som är inflytelserika.

3.2.4 Modellval

Det kan vara så att vissa variabler inte förklarar responsvariabeln något nämnvärt, dessa sägs vara statistiskt insignifikanta. Ett sätt att avgöra huruvida en variabel är signifikant eller ej är genom att undersöka dess p värde. Man testar då nollhypotesen att variabeln ej påverkar responsvariabeln mot mothypotesen att variabeln påverkar responsvariabeln. Ifall en variabel är insignifikant bör den tas bort eftersom det skulle ge en enklare modell utan nämnvärd förlust.

Det finns anledningar att vilja förenkla modellen även om alla variabler är signifikanta, kom ihåg Einstains uttryck i avsnitt 3.1. Det kan vara så att ett litet urval av de förklarande variablerna förklarar majoriteten av den totala variansen och att det i sådana fall skulle vara bättre att använda den enklare modellen. Detta kan undersökas med hjälp av stegvis variabelselektion vilket går ut på att man succesivt lägger till eller tar bort en variabel åt gången tills ett stoppkriterium är uppfyllt, exempeltvis AIC.

AIC ($Akaike\ Information\ Criterion$) används för att finna den modell som förklarar responsvariabeln bäst med minst antal förklarande variabler. AIC beräknas enligt:

$$AIC = 2k - 2ln(\widehat{L}) \tag{5}$$

där k är antalet parametrar och $\widehat{L}$ är den maximerade likelihood funktionen.

3.2.5 Multikollineära variabler

Multikollineära variabler kan, som nämndes i avsnitt 3.1, orsaka flertalet problem. Ett av dessa är att två eller fler multikollineära variabler kan bli insignifikanta tillsammans men signifikanta enskilda. Ett mått som kan användas för att undersöka multikollinearitet är VIF ($Variation\ Inflation\ Factor$). VIF kan beräknas enligt:

$$VIF = \frac{1}{1 - R_j^2} \tag{6}$$

där R_j^2 står för hur mycket av variationen i x_j som förklaras av de övriga x variablerna.

3.3 Principalkomponent regression

Följande avsnitt om PCR är baserad på kapitel 6.3, 6.3.1 och 10.2 i *An Introduction To Statistical Learning* [7]. Mer specifikt beskrivs principalkomponentanalys i kapitel 10.2 medan principalkomponent regression beskrivs i kapitel 6.3 och 6.3.1.

Principalkomponent regression är en regression som lämpar sig väl på datamaterial med många förklarande variabler och/ eller där det finns multikollineära variabler. I principalkomponent regression skapar man principalkomponenter vilka sedan används som förklarande variabler i en (multipel) linjär regressionsmodell. Resultatet av detta är att majoriteten av de urspungliga variablernas varians kan förklaras med hjälp av färre antal okorrelerade variabler. Denna metod minskar således risken för att modellen överanpassas samt eliminerar risken för multikollineära variabler.

3.3.1 Principalkomponentanalys

Principalkomponentanalys är processen att skapa principalkomponenter. Den första principalkomponenten av de centrerade, förklarande variablerna $X_1, ..., X_p$ är linjärkombinationen

$$Z_1 = \phi_{11} \cdot X_1 + ... + \phi_{p1} \cdot X_p \tag{7}$$

som har högst varians och uppfyller $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Vi hittar den första principalkomponenten genom att hitta den linjärkombination

$$z_{i1} = \phi_{11} \cdot x_{i1} + ... + \phi_{p1} \cdot x_{ip} \tag{8}$$

som maximerar stickprovsvariansen där $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Elementen $z_{11}, ..., z_{n1}$ kallas den första principalkomponentens *scores* och $\phi_{11}, ..., \phi_{p1}$ kallas den första principalkomponentens *loadings*. Stickprovsvariansen kan uttryckas $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$ eftersom $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ vilket medför att $\frac{1}{n}\sum_{i=1}^{n} z_{i1} = 0$. Den första principalkomponentens *loading* vektor maximerar således

$$\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p} \phi_{j1} x_{ij})^2 \tag{9}$$

där $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. *Loading* vektorn som maximerar (9) kan man finna genom singulärvärdesuppdelning. Detaljerna kring singulärvärdesuppdelningen kan man läsa om på *slide* 8 i [8]. Singulärvärdesuppdelningen ger egenvärden där egenvektorn för det största egenvärdet är detsamma som den första principalkomponentens *loading* vektor.

Den andra principalkomponenten Z_2 är den linjärkombination av $X_1, ..., X_p$ som har högst varians av de linjärkombinationer som är okorrelerade med Z_1. Resterande principalkomponenter skapas på samma sätt, totalt skapas p principalkomponenter.

Innan principalkomponenter skapas ska de förklarande variablerna somsagt

centreras till att ha väntevärde 0 och de bör standardiseras till att ha standardavvikalse 1. Ifall man inte standardiserar variablerna kommer de variabler med hög varians väga tyngre i skapandet av principal komponenterna än de med låg varians. Variablerna standardiseras genom följande formel:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_{ij})^2}}. \tag{10}$$

3.3.2 PVE

I principalkomponentanalys är det viktigt att utvärdera hur mycket varians respektive principalkomponent kan förklara för att kunna välja rätt mängd principalkomponenter. Man vill alltså veta PVE (*Proportion of Variance Explained*). Den totala variansen, givet att de förklarande variablernas har centrerats, ges av:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2. \tag{11}$$

Variansen för principalkomponent m ges av:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p} \sigma_{jm} x_{ij})^2. \tag{12}$$

PVE för principalkomponent m ges av:

$$\frac{\sum_{i=1}^{n}(\sum_{j=1}^{p} \sigma_{jm} x_{ij})^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2}. \tag{13}$$

Den kumulativa PVE för de m första principalkomponenterna erhålls genom att summera (13) för respektive principalkomponent.

3.3.3 Välja antalet principalkomponenter

Principalkomponentanalys används dels för att reducera datamaterialets dimension vilket betyder att vi vill välja så få principalkomponenter som möjligt som förklarar så mycket av variansen som möjligt. Man kan välja antalet principalkomponenter genom att plotta PVE samt den kumulativa PVE mot antalet principalkomponenter. Man vill då hitta den så kallade armbågs punkten i plottarna.

3.3.4 Faktor analys av blandad data

En nackdel med PCA är att PCA inte kan appliceras på kategoriska variabler. Det finns dock en teknik, som kallas $FAMD$ (*Factor Analysis of*

Mixed Data), som kringgår denna nackdel. $FAMD$ är en förlängning av PCA som kan anta både numeriska och kategoriska variabler. Detta avnitt om FAMD är baserad på en artikel från towardsdatascience.com [9].

Till att börja med ska vi definera principalkomponenterna på ett annat sätt. Principalkomponenterna defineras nu som den centrerade, enhetsvektorn som är ortogonal mot resterande principalkomponenter samtidigt som den maximerar den projicerade *inertian* för variablerna v_j. *Inertian* defineras som:

$$I = \sum_{j=1}^{P} p_j ||v_j||^2 \tag{14}$$

där v_j är kolumnvektorn för de förklarande variablerna, $||v_j||^2$ är normen av v_j och p_j är vikten på respektive variabel. I PCA sätter man oftast $p_j = 1$ för att låta alla variabler påverka *inertian* lika mycket. Den projicerade *inertian* för variablerna längs en vektor w defineras som:

$$I_{/w} = \sum_{j=1}^{P} p_j < v_j, \frac{w}{||w||} >^2 \tag{15}$$

där $< v_j, \frac{w}{||w||} >=< v_j, w >= \frac{1}{n}\sum_{i=1}^{n}(v_j)_i w_i$ eftersom $||w|| = 1$. I PCA väljs som sagt den första principalkomponenten som den vektor som maximerar den projicerade inertian. Nackdelen med PCA är att det ej kan appliceras på kategoriska variabler. Man skulle kunna omvandla de kategoriska variablerna till dummy variabler men då skulle den projicerade inertian bero på antalet klasser samt sannolikheten att de antas i de kategoriska variablerna. Detta tillvägagångssätt gör det således svårt att ge de numeriska och kategoriska variablerna samma betydelse i skapandet av principalkomponenterna. Idèn med $FAMD$ är att man kan skriva om den projicerade *inertian* så att den kan uttryckas av korrelationen mellan v_j och w. Anledningen till att man vill skriva om den projicerade *inertian* på detta vis är för att korrelationen kan uttryckas inom samma intervall för både numeriska och kategoriska variabler. På så sätt kan de numeriska och kategoriska variabler ge samma betydelse i skapandet av principalkomponenterna.

$FAMD$ bygger som sagt på en idè där man skriver om den projicerade inertian så att den beror på korrelationen mellan v_j och w. Detta uppnås genom att standardisera samt centrera de numeriska variablerna medan de kategoriska variablerna omvandlas till dummy variabler som sedan divideras med sannolikheten att klassen antas, μ_j, samt centreras. Sedan ges vikten $p_j = 1$ till de numeriska variablerna och vikten $p_j = \mu_j$ till dummy variablerna. Resultatet av detta är att man kan skriva om den projicerade *inertian*. Man kan läsa om omskrivningsstegen i avsnitt $V]Demonstrations$ i artikel

[9] men här nöjer vi oss med att konstatera att omskrivningen resulterar i den fundamentala ekvationen för $FAMD$. Den fundamentala ekvationen för $FAMD$ är:

$$I_{/w} = r^2(v_j, w) - numerisk, \tag{16}$$

$$I_{/w} = \eta^2(v_j, w) - kategorisk,$$

där $r^2(v_j, w)$ är Pearsons korrelationskoefficient i kvadrat och $\eta^2(v_j, w)$ är korrelationsförhållandet mellan en kategorisk och en numeriska variabel. Både Pearsons korrelationskoefficient i kvadrat och korrelationsförhållandet antar värden inom intervallet $[0, 1]$ där högre värde innebär högre korrelation. Vi kan nu definiera principalkomponenterna för ett datamaterial med P numeriska variabler och Q kategoriska variabler som

$$Z_k = \underset{||w||=1, \overline{w}=0}{argmax} \{\sum_{j=1}^{P} r^2(w, v_j) + \sum_{j=P+1}^{P+Q} \eta^2(w, v_j | w \perp Z_1, ...Z_{k-1}).\} \tag{17}$$

Den första principalkomponenten hittas således genom att hitta den centrerade enhetsvektorn w som maximerar $r^2(v_j, w)$ när v_j är numerisk och $\eta^2(v_j, w)$ när v_j är kategorisk.

3.4 Generaliserad additiv modell

Följande avsnitt om den generaliserade additiva modellen är baserat på kapitel 7.1 - 7.5 och 7.7 - 7.7.1 i *An Introduction To Statistical Learning* i [7].

En generaliserad additiv modell (GAM) är lik den multipla linjära regressionsmodellen i form av addition mellan variablerna men skiljer sig i form av att GAM tillåter icke linjär funktioner på de förklarande variablerna. Innan vi går in på detaljerna kring GAM ska vi gå igenom några olika funktioner och metoder.

3.4.1 Steg funktioner

Steg funktioner delar en variabels intervall i K olika regioner och producerar kategoriska variabler. Låt oss tänka oss en variabel X som antar värden inom intervallet $[c_0, c_{K+1}]$. En steg funktion skapar regionsgränserna $c_1, ..., c_K$ och sedan $K + 1$ nya variabler av formen:

$$C_0(X) = I(X < c_1),$$

$$C_1(X) = I(c_1 \leq X < c_2),$$

$$...$$

$$C_K(X) = I(c_K \leq X). \tag{18}$$

$I(\cdot)$ är en indikator funktion som ger 1 ifall det stämmer och 0 annars. Sedan används minsta kvadrat metoden för att anpassa den linjära modellen

$$y_i = \beta_0 + \beta_1 C_1(x_1) + ... + \beta_K C_K(x_K) + \epsilon_i \tag{19}$$

där β_j representerar medelvärdet av ökningen i responsvariabeln för $c_j \leq X < c_{j+1}$ relativt till $X < c_1$.

3.4.2 Polynomregression

Polynomregression är en förlängning av den linjära regressionsmodellen där man adderat förklarande variabler. Modellen för polynomregression skrivs:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ...\beta_d x_i^d + \epsilon_i. \tag{20}$$

3.4.3 Bas funktioner

Polynomregressionsmodeller och regressionsmodeller med steg funktioner är modeller med bas funktioner. Istället för att anpassa en linjär modell med förklarande variabel X anpassar man följande modell:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + ... + \beta_K b_K(x_i) + \epsilon_i \tag{21}$$

där $b_j(\cdot)$ är bas funktionerna. Bas funktionerna för polynomregression är $b_j(x_i) = x_i^j$ och för steg funktioner har vi $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$. Nu ska vi gå in på regression *splines* vilket är vanliga bas funktioner.

3.4.4 Regression splines

Regression *splines* är en kombination av steg funktioner och polynomregression. I regression *splines* delar man intervallet av en variabel i K olika regioner och anpassar ett polynom inom varje region. Regression *splines* liknar styckvis polynomregression men i regression *splines* har polynomen begränsningen att de ska vara kontinuerliga vid regionsgränserna, även kallat knutarna. Man kan även ansätta fler begränsningar på polynomen ifall de är för flexibla. Man kan exemepltvis kräva att polynomen ska vara kontinuerlig och smooth vid knutarna genom att kräva att polynomen samt första och andra derivatan av dem ska vara kontinuerliga vid knutarna. Naturliga *splines* är regression *splines* med gränsbegränsningarna att funktionen ska vara linjär vid gränserna.

Man kan som sagt använda en modell med bas funktioner för att representera regression *splines*. Låt oss kolla på hur man modellerar kubiska *splines* med hjälp av bas funktioner som ett exempel. Grad d *splines* är styckvisa grad d polynom med kontinuerlig derivata upp till grad $d - 1$ i varje knut. Kubiska *splines* är således styckvisa grad 3 polynom med kontinuerlig första

och andra derivata i varje knut. Kubiska *splines* med K knutar kan, med bas funktionerna $b_1, ..., b_{K+3}$, skrivas

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i. \tag{22}$$

I modell (22) väljer man bas funktionerna till $b_1(x_i) = x, b_2(x_i) = x^2, b_3(x_i) = x^3$ och $b_k(x_i) = h(x, \zeta_k)$ där $h(x, \zeta_k)$ är en *truncated power* bas för knuten ζ_k. En *truncated power* bas definieras som

$$h(x, \zeta)^3 = (x - \zeta)^3 \text{ om } x > \zeta \text{ och } 0 \text{ annars} \tag{23}$$

där ζ är knuten. En modell med kubiska splines och K knutar kan anpassas med minsta kvadratmetoden på intercept och de 3 + K förklarande variablerna $X, X^2, X^3, h(X, \zeta_1), ..., h(X, \zeta_K)$. Detta ger således $4 + K$ parametrar som ska skattas vilket betyder att kubiska splines har $4 + K$ frihetsgrader. Mer utförlig beskrivning av hur man representerar regressions splines med bas funktioner finns i kapitel 7.4.3 i [7].

Regression *splines* är mest flexibla i regioner med mycket knutar eftersom koefficienterna där ändras snabbt. I regression *splines* måste man välja antalet knutar samt vart dessa ska placeras. Vi såg i exemplet med kubiska *splines* att man kan avgöra hur många knutar en regression *spline* har om man vet antalet frihetsgrader modellen har, detsamma gäller även åt andra hållet. I praktiken kan man således välja antalet knutar genom att, med korsvalidering, hitta antalet frihetsgrader som resulterar i lägst $MSEP$ och sedan låta programmet beräkna motsvarande antal knutar och placera dessa jämt över variabelns kvantiler.

3.4.5 Smooth spline

När man anpassar en modell med *smooth splines* vill man hitta en funktion $g(x)$ som anpassar datamaterialet bra och är *smooth*. En sådan funktion hittar man genom att hitta funktionen g som minimerar

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt \tag{24}$$

där λ är en icke negativ *turning* parameter. Funktionen som minimerar (24) kallas *smooth spline*, det visar sig även att denna funktion är en naturlig kubisk *spline* med knutar i varje x_i.

Man kan tro att *smooth splines* får för många frihetsgrader eftersom varje knut ger mer flexibilitet men i *smooth splines* mäts flexibiliteten av effektiva frihetsgrader vilket kontrolleras av λ. Allteftersom λ blir större blir g mer *smooth* och antalet effektiva frihetsgrader minskar. Vi kan skriva $g_\lambda = S_\lambda y$

där g_λ är lösningen för (24) given en λ och y är respons vektorn. De effektiva frihetsgraderna kan nu definieras som

$$df_\lambda = \sum_{i=1}^{n} \{S_\lambda\}_{ii}. \tag{25}$$

Man kan läsa mer om de effektiva frihetsgraderna i kapitel 7.5.2 i [7].

Smooth splines har somsagt knutar vid varje x_j, vi behöver således ej välja antalet knutar eller platserna för dessa. Istället måste man välja λ. I praktiken kan man välja λ genom att, med korsvalidering, hitta antalet effektiva frihetsgrader som resulterar i lägst $MSEP$ och sedan låta programmet beräkna motsvarande λ.

3.4.6 GAM

Vi har nu gått igenom hur man kan anpassa en modell med bas funktioner för en förklarande variabel. Nu ska vi gå igenom hur man kan anpassa en modell med bas funktioner för flera förklarande variabler. Detta kan göras med *GAM* (*Generalized Additive Model*) som är lik en multipel linjär regression i form av addition mellan variablerna men olik i form av att icke linjära funktioner tillåts på variablerna. Vi kan skriva *GAM* genom att byta ut $\beta_j x_{ij}$ i en multipel linjär regression mot en icke linjär funktion $f_j(x_{ij})$. Modellen skrivs

$$y_i = \beta_0 + f_1(x_{i1}) + ... + f_p(x_{ip}) + \epsilon_i. \tag{26}$$

Det som är bra med *GAM* är att man bland annat kan använda metoderna vi gått igenom ovan på variablerna för att anpassa en additiv modell. I detta arbete kommer vi testa naturliga samt *smooth splines* på de numeriska variablerna och steg funktioner på de kategoriska variablerna. Fördelen med *GAM* med naturliga *splines* är att modellen kan anpassas med minsta kvadrat metoden. *GAM* med *smooth splines* kan dock inte anpassas med minsta kvadrat metoden, där anpassar man istället modellen med något som kallas *backfitting*. *Backfitting* anpassar en modell med flera förklarande variabler genom att upprepande gånger uppdatera anpassningen för varje förklarande variabel medan resterande variabler hålls fixerade.

3.5 Prediktering

När modellerna är klara kommer det vara av intresse att utvärdera deras prediktionsförmågor. Det finns ett flertal sätt att prediktera observationer på. Man kan använda hela datamaterialet för att anpassa modellen och göra predikteringar. Risken med detta tillvägagångssätt är att man kan få missvisande bra predikteringar eftersom man gör predikteringar på data

som anpassat modellen. Man kan undvika detta problem genom att dela datamaterialet i träningsdata och testdata och använda träningsdatan till att anpassa modellen medan testdatan används för att göra predikteringar. Nackdelen med detta tillvägagångssätt är att man går miste om information i anpassningen och predikteringen. I detta arbete kommer istället ett annat tillvägagångssätt användas, nämnligen *LOOCV* (*Leave One Out Cross Validation*). Anledningen till att denna metod väljs är för att den delar datamaterialet så att predikteringar kan göras på observationer som ej var delaktiga i anpassningen av modellen samtidigt som den bevarar så mycket data som möjligt i anpassningen och predikteringarna.

LOOCV beskrivs på sida 178 i *An Introduction To Statistical Learning* [7] och innebär att man delar datamaterialet i två delar, valideringsdata samt träningsdata. I *LOOCV* utgörs valideringsdatan av en observation (x_1, y_1) och träningsdatan av resterande observationer $(x_2, y_2), ..., (x_n, y_n)$. Modellen anpassas därefter till $(x_2, y_2), ..., (x_n, y_n)$ och en prediktering görs för den exkluderade observationen med hjälp av x_1. Denna procedur görs för alla observationer.

3.5.1 Prediktionsmått

När man gjort predikteringar, vilket vi kommer göra med *LOOCV*, kommer det vara av intresse att undersöka hur väl modellen predikterat ny data. Detta kan beskrivas med prediktionsmåttet *RMSEP* (*Root Mean Squared Error Of Prediction*). *RMSEP* definieras enligt:

$$RMSEP = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \widehat{y_i})^2}{n}} \tag{27}$$

där y_i är det observerade värdet på observation i och $\widehat{y_i}$ är det predikterade värdet på observation i då modellen anpassats utan observationen. Man kan även beskriva prediktionsförmågan med *MAEP* (*Mean Absolute Error of Prediction*). *MAEP* definieras enligt:

$$MAEP = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n} \tag{28}$$

där y_i och $\widehat{y_i}$ är samma som i *RMSEP*.

4 Statistisk modellering

4.1 Multiple linjär regression

Det är nu dags att ställa upp en multipel linjär regressionsmodell för att beskriva IMDb-betyget. Innan modellen ställs upp undersöks de variabler som tidigare visade sig vara multikollineära.

4.1.1 Multikollineära variabler

Det visade sig i avsnitt 2.4 att **Avsnittsnummer**, **Säsong** och **Avsnitt** är multikollinära. Multikollinearitet kan som sagt göra så att signifikanta variabler blir insignifikanta tillsammans, det är därmed viktigt att åtgärda multikollineära variabler. De tre multikollineära variablerna beskriver mycket liknande information så det beslutas att exkludera någon av dem istället för att byta ut dem mot en ny variabel.

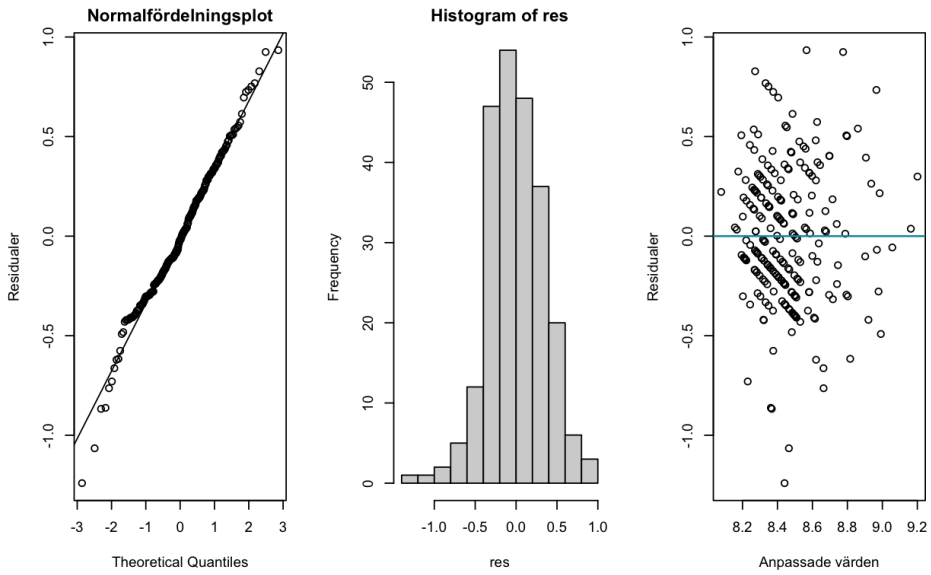
Kom ihåg att **Säsong** och **Avsnittsnummer** är starkt parvist korrelerade. Åtminstone en av dessa variabler behöver således exkluderas för att bli av med mulitkollinearitеten. Vi beslutar att exkludera **Avsnittsnummer** eftersom arbetet handlar om en serie och ett avsnitts placering i serien lämpligen beskrivs av dess säsong samt avsnittsordning inom säsongen. Resultatet av att exkludera **Avsnittsnummer** är att VIF värdena för variablerna ligger inom intervallet $[1.115, 2.228]$, det finns således ingen multikollinearitet kvar.

4.1.2 Första multipla linjära regressionsmodellen

Den första multipla linjära regressionsmodellen som undersöks är en modell som innehåller alla förklarande variabler, bortsett från **Avsnittsnummer**, samt alla observationer. Modellen har således 25 förklarande variabler och 236 observationer. Modellen skrivs:

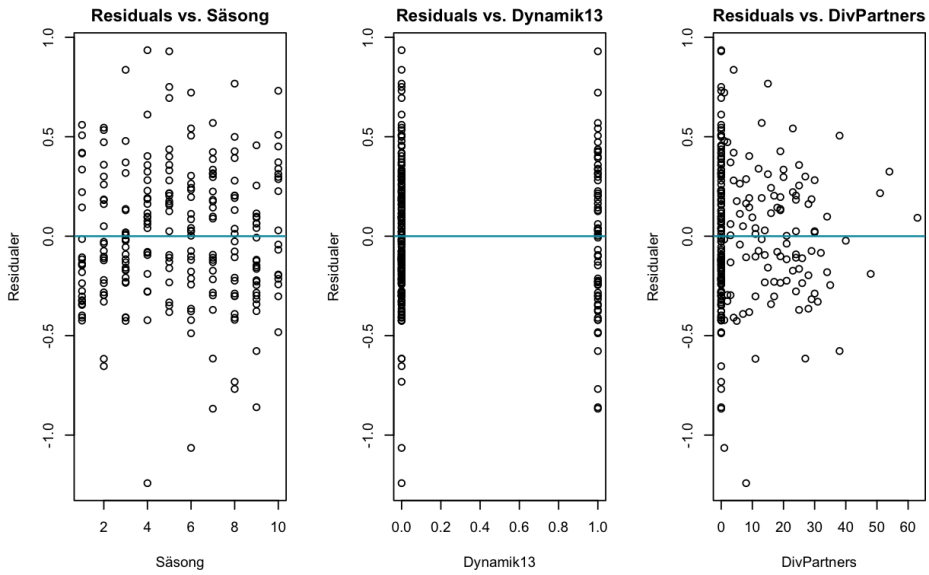
$$IMDb = \alpha + \beta_1 \cdot Säsong + \beta_2 \cdot Avsnitt + ... + \beta_{25} \cdot Övriga + \epsilon \qquad (29)$$

Residualplottarna för modell (29) syns nedan.



Figur 3: Residualplottar för modell (29)

I normalfördelningsplotten i figur 3 syns avikelser från normalfördelningen för främst låga kvantiler. Histogrammet i figur 3 visar en vänstersked normalfördelning. I den högra plotten i figur 3 syns ett diagonaliskt mönster vilket antyder att feltermerna inte är oberoende vilket kan betyda att en linjär modell inte är att föredra. I figurerna nedan syns residualerna plottade mot ett urval av de förklarande variablerna.



Figur 4: Residualer plottade mot ett urval av de förklarande variablerna

När man plottar residualerna mot de förklarande variablerna ger dummy variablerna ett väntat mönster likt den mellersta plotten i figur 4. Variablerna som har många observationer som svarar mot ett specifikt värde ger plottar med mönster liknande mönstret som ses i den högre plotten i figur 4. Utöver dessa variabler ger **Säsong** ett mönster enligt den vänstra plotten i figur 4, resterande variabler ger plottar utan synligt mönster. Det diagonaliska mönstret i den högra plotten i figur 3 kan således bero på att sambandet mellan de förklarande variablerna och responsvariabeln inte är linjärt. Vi kommer därför, senare i arbetet, även testa en modell som tillåter icke linjära funktioner på variablerna, nämnligen GAM.

Modell (29) har $R^2 = 0.245$, $R^2_{adj} = 0.155$ samt $p - värde = 5.492 \cdot 10^{-05}$. Residualerna för en modell kan ibland bli bättre med en tranformation på responsvariabeln. Transformationerna $log(IMDb)$, $\frac{1}{IMDb}$ samt $\sqrt{IMDb}$ testas men förbättrar inte residualerna. Nästa steg blir nu att undersöka ifall det finns några avvikande observationer som bör exkluderas.

4.1.3 Avvikande observationer

De avvikande observationerna identifieras med Cooks avstånd. Kriterierna $\frac{4}{N-k-1}$, $\frac{4}{N}$ och $3 \cdot mean(D_i)$ testas där $3 \cdot mean(D_i)$ ger bäst resultat i form av $RMSEP$, R^2 och $MAEP$. Cooks avstånd med detta kriterium fann 20 avvikande observationer. Något tydligt samband mellan alla dessa observationer syns ej i datamaterialet men det finns en del samband mellan vissa av dessa

observationer. Det visar sig att 6 av de lägsta samt 3 av de högsta observerade IMDb-betygen tillhör de avvikande observationerna. Av dessa följer majoriteten av avsnitten en historia centrerad runt Chandler och Monica eller Rachel och Ross. I de resterande 11 observationerna följer majoriteten av avsnitten en historia centrerad runt en huvudkaraktärs familjemedlem. Frågan är nu huruvida de avvikande observationerna ska inkluderas i modellen eller ej. Avvikande observationer kan innehålla viktig information vilket anses vara fallet med de identifierade observationerna. I slutet av detta arbete ska även de olika modellerna jämföras vilket kommer vara svårt ifall de olika modellerna innehåller olika observationer. Beslutet tas därmed att behålla alla 236 observationer.

4.1.4 Modellval

Modellen innehåller nu alla variabler utom **Avsnittsnummer** samt alla observationer. Nästa steg i modelleringen är att undersöka vilka variabler som ska inkluderas i den slutgiltiga modellen. Detta undersöks med hjälp av stegvis variabelselektion åt båda håll där AIC används som stoppkriterium. Resultatet av detta är en modell med de förklarande variablerna **KevinR**, **MichaelR**, **AlexaM**, **Längd**, **Thanksgiving**, **Dynamik56**, **Dynamik13** och **RossGeller**. Modellen har $AIC = -479$. Det visar sig dock att dummy variabeln som visar huruvida manuset skrivits av Alexa har $p - värde = 0.131$, dummy variabeln som visar huruvida avsnittet följer en historia om Rachel och Ross har $p - värde = 0.121$ och dummy variabeln som visar huruvida Michael regisserat avsnittet har $p - värde = 0.119$. Dessa tre variabler är således insignifikanta på 5% signifikansnivå. Frågan är då huruvida de insignifikanta variablerna ska inkluderas i modellen eller ej. En modell med de insignifikanta variablerna ger ett något lägre AIC värde, $AIC = -479$ jämfört med $AIC = -477$. Modellen med de insignifikanta variablerna ger dessutom betydligt mycket bättre residualer samt bättre anpassningsmått, $R^2 = 0.222$ i jämförelse med $R^2 = 0.195$ samt $R^2_{adj} = 0.195$ i jämförelse med $R^2_{adj} = 0.177$. Det beslutas därmed att välja en 15% signifikansnivå vilket betyder att vi väljer modellen som gavs av stegvis variabel selektion.

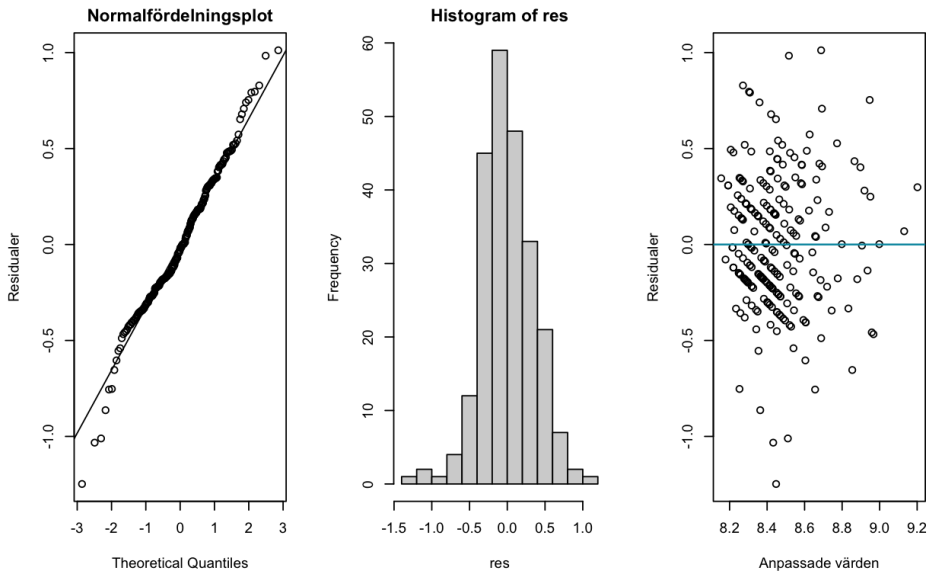
4.1.5 Andra multipla linjära regressionsmodell

Den andra och slutgiltiga multipla linjära regressionsmodellen har de förklarande variabler **KevinR**, **MichaelR**, **AlexaM**, **Längd**, **Thanksgiving**, **Dynamik56**, **Dynamik13** och **RossGeller** samt alla observationer. Modellen skrivs:

$$IMDb = Intercept + \beta_1 \cdot KevinR + \beta_2 \cdot MichaelR + \beta_3 \cdot AlexaM + \beta_4 \cdot Längd + \beta_5 \cdot Thanksgiving$$

$$+ \beta_6 \cdot Dynamik56 + \beta_7 \cdot Dynamik13 + \beta_8 \cdot RossGeller + \epsilon \qquad (30)$$

Modell (30) har $R^2 = 0.222$, $R^2_{adj} = 0.195$ samt $p - värde = 1.317 \cdot 10^{-9}$ och dess residualplottar syns nedan i figur 5.



Figur 5: Residualplottar för modell (30)

Den slutgiltiga multipla linjära regressionsmodellens VIF värden, parameterskattningar samt $p - värden$ finns nedan i tabell 1.

Tabell 1: VIF värden, parameterskattning samt p-värde för modell (30)

Variabel	VIF	Parameterskattning	P-värde
Intercept	-	7.082	-
KevinR	1.200	0.161	0.008
MichaelR	1.056	0.123	0.119
AlexaM	1.028	0.169	0.131
Längd	1.082	0.047	0.003
Thanksgiving	1.035	0.398	0.001
Dynamik56	1.152	0.085	0.120
Dynamik13	1.132	0.129	0.021
RossGeller	1.141	0.005	0.009

Till sist ska modellens prediktionsförmåga utvärderas. Prediktionsförmågan utvärderas med LOOCV och resultatet syns i tabellen nedan. Här har R^2 erhållits genom att beräkna Pearsons korrelationskoefficient mellan det observerade och predikterade IMDb-betyget i kvadrat.

Tabell 2: Prediktionsförmåga med LOOCV för modell (30)

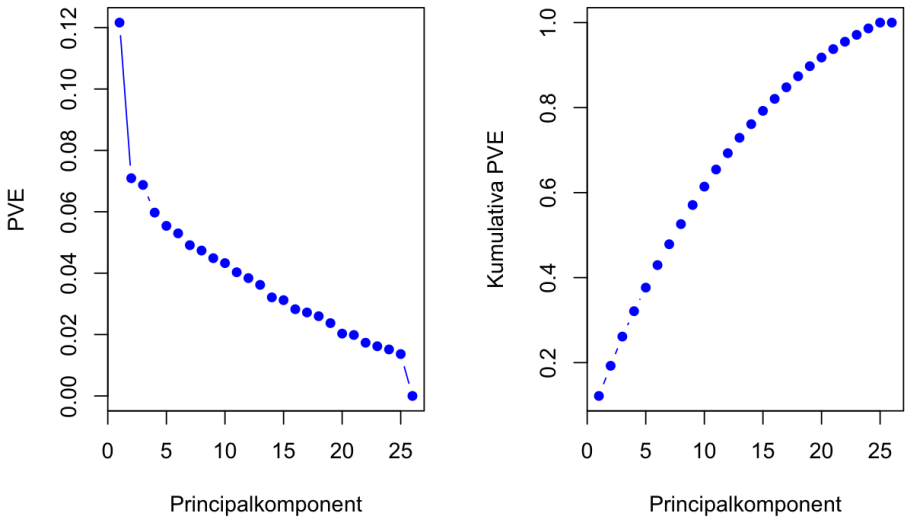
	R^2	RMSEP	MAEP
Modell (30)	0.147	0.367	0.289

4.2 Principalkomponent regression (PCR)

Vi ska nu testa en annan regressionsmodell, nämnligen principalkomponent regression.

4.2.1 PCR på hela datamaterialet

Den första principalkomponent regressionsmodellen som ska testas är en modell där alla 26 förklarande variablerna är med i skapandet av principalkomponenterna. Vi skapar de 26 okorrelerade principalkomponenterna med hjälp av R:s inbyggda funktioner. Vi ska nu undersöka hur mycket varians som kan förklaras beroende på antalet principalkomponenter som väljs. Detta undersöks med PVE (*Proportion of Variance Explained*) och görs för att kunna välja antalet principalkomponenter att inkludera i den multipla linjära regressionsmodellen.



Figur 6: PVE och kumulativa PVE för principalkomponenterna

Plottarna i figur 6 undersöks för att bestämma antalet principalkomponenter som ska användas i den multipla linjära regressionsmodellen. Den sökta punkten i den vänstra plotten i figur 6 är den så kallade armbågspunkten.

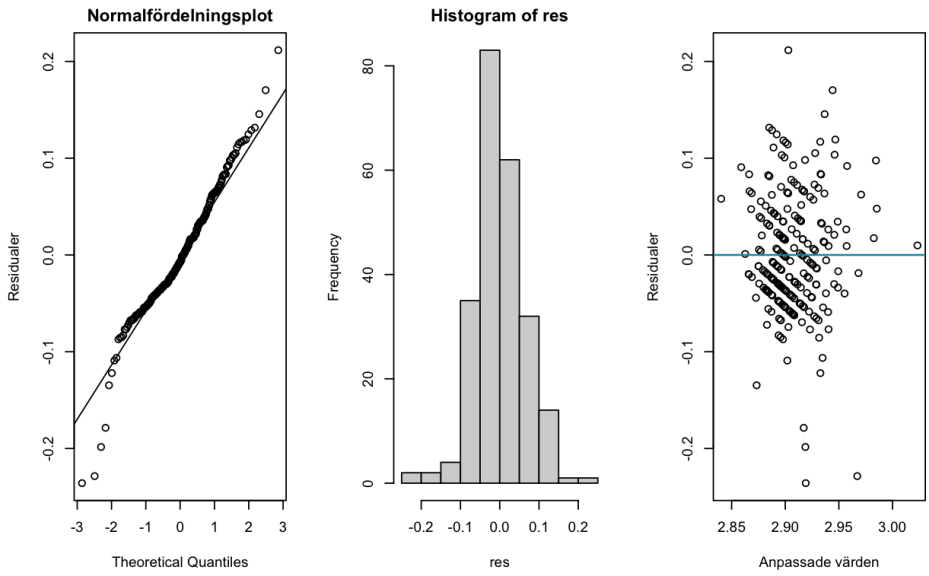
Armbågspunkten sker vid principalkomponent 2 och 4. Man kan argumentera för att principalkomponent 26 även utgör en armbågspunkt men då målet med PCR är att minska dimensionen i datamaterialet kommer modellen ställas upp med två respektive fyra principalkomponenter. Första PCR modellen som ska ställas upp är en multipel linjär regressionsmodell med de två första principalkomponenterna. Modellen skrivs:

$$IMDb_i = \alpha + \beta_1 \cdot PC1_i + \beta_2 \cdot PC2_i + \epsilon_i. \tag{31}$$

Modell (31) ger $R^2 = 0.139$, $R^2_{adj} = 0.132$, $p - värde = 2.624 \cdot 10^{-08}$ samt residualer som uppfyller kraven dåligt. Det testas några olika transformationer där responsvariabel $\sqrt{IMDb}$ visar sig vara den bästa. Modellen med responsvariabel $\sqrt{IMDb}$ kan skrivas:

$$\sqrt{IMDb_i} = \alpha + \beta_1 \cdot PC1_i + \beta_2 \cdot PC2_i + \epsilon_i. \tag{32}$$

Residualplottarna för modell (32) finns nedan i figur 7.



Figur 7: Residualplottar för modell (32)

Modell (32) har $R^2 = 0.136$, $R^2_{adj} = 0.129$, $p - värde = 4.046 \cdot 10^{-08}$ samt residualer som uppfyller kraven bättre än modell (31). Prediktionsförmågan testas med $LOOCV$. Resultatet av $LOOCV$ ges i tabellen nedan.

Tabell 3: Prediktionsförmåga med LOOCV för modell (32)

	R^2	RMSPE	MAEP
Modell (32) med $\sqrt{IMDb}$	0.117	0.064	0.049
Modell (32) med IMDb	0.117	0.372	0.287

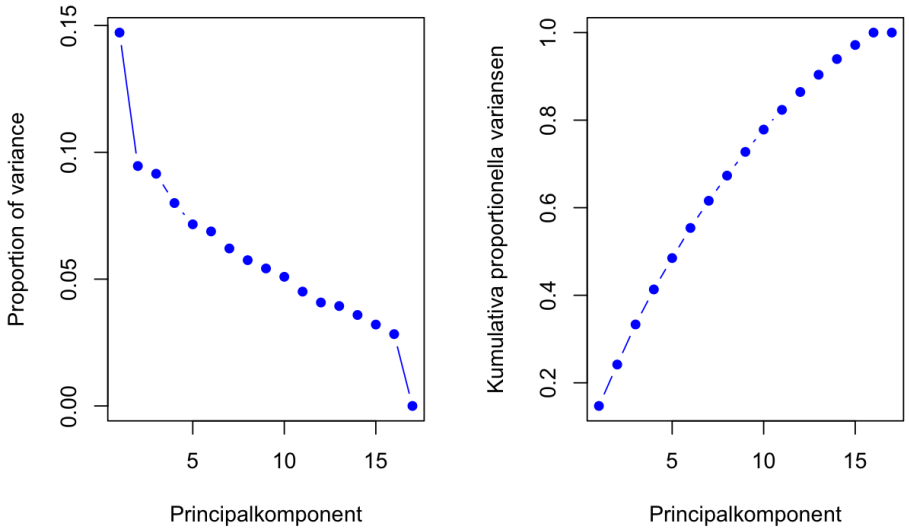
Rad 2 i tabell 3 anger värden som beräknats med responsvariabel $\sqrt{IMDb}$. I rad 3 i tabell 3 har värdena istället beräknats på $IMDb$ genom att ta det predikterade värdet i kvadrat. Värdena i rad 3 i tabell 3 har beräknats för att enklare kunna jämföra de olika modellerna.

En modell med fyra principalkomponenter testas också men ger likvärdigt resultat som modell (32). Det eftersträvas alltid att ha så enkla modeller som möjligt så vi konstaterar att det är bättre att modellera PCR med två principalkomponenter än med fyra i detta fall.

Vi undersöker nu ifall det går att erhålla en bättre PCR modell än den med alla variabler. Det nämndes i avsnitt 3.2.4 att det kan vara olämpligt att applicera PCA på datamaterial med dummy variabler. Vi testar därmed, i nästa avsnitt, en modell där princiaplkomponenterna skapas utan dummy variablerna. Dummy variablerna läggs sedan till som de är i den multipla linjära regressionsmodellen.

4.2.2 PCA på datamaterialet utan dummy variablerna

Vi genomför nu PCA utan dummy variablerna **GaryR**, **KevinR**, **MichaelR**, **AlexaM**, **AndrewTedM**, **Thanksgiving**, **Dynamik13**, **DYnamik56** och **Dynamik12**. PCA genomförs således på 17 variabler. Samma steg som i föregående avsnitt genomförs och nedan syns PVE samt kumulativa PVE plottarna för de 17 skapade principalkomponenterna.

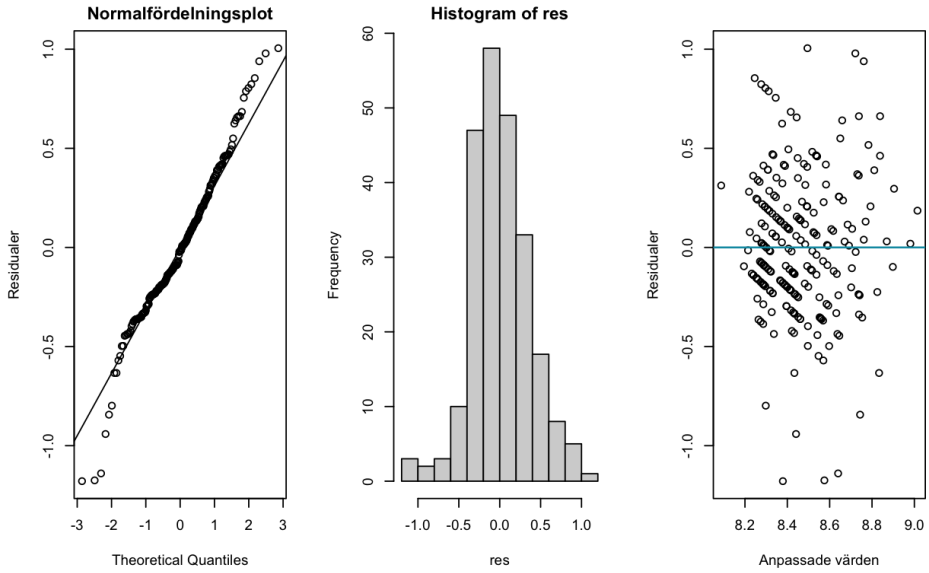


Figur 8: PVE och kumulativa PVE för principalkomponenterna

Den vänsta plotten i figur 8 visar att armbågspunkterna sker vid 2, 4, 5 respektive 17 principalkomponenter. PCR är somsagt en dimensionsreduceringsteknik så vi väljer att ställa upp linjära regressionsmodeller med 2, 4 och 5 principalkomponenter som förklarande variabler. Först ut testas en multipel linjär regressionsmodell med 2 principalkomponenter och de 9 dummy variablerna som förklarande variabler. Modellen skrivs:

$$IMDb_i = \alpha + \beta_1 \cdot PC1_i + \beta_2 \cdot PC2_i + ... + \beta_{11} \cdot Dynamik13 + \epsilon_i \quad (33)$$

Modell (33) har $R^2 = 0.187$, $R^2_{adj} = 0.147$ samt $p - värde = 1.875 \cdot 10^{-6}$. Residualplottarna för modellen syns nedan.



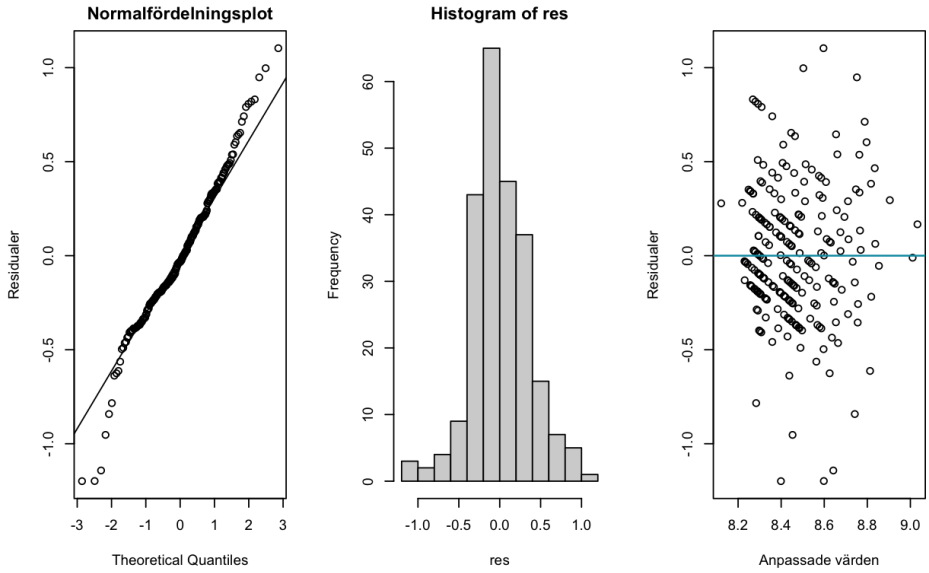
Figur 9: Residualplottar för modell (33)

Det testas ifall någon transformation på responsvariabeln förbättrar residualerna men ingen sådan transformation hittas. Nästa steg blir nu att undersöka vilka variabler som ska inkluderas genom stegvis variabelselektion. Resultatet av detta är en modell med de 7 förklarande variablerna **PC2**, **KevinR**, **AlexaM**, **Thanksgiving**, **Dynamik56**, **Dynamik12** och **Dynamik13**. Modellen skrivs

$$IMDb = \alpha + \beta_1 \cdot PC2 + \beta_2 \cdot KevinR + \beta_3 \cdot AlexaM + \beta_4 \cdot Thanksgiving$$

$$+\beta_5 \cdot Dynamik56 + \beta_6 \cdot Dynamik12 + \beta_7 \cdot Dynamik13 + \epsilon. \qquad (34)$$

Residualplottarna för modell (34) finns nedan.



Figur 10: Residualplottar för modell (34)

Figur 10 visar att modell (34) har residualer som uppfyller kraven någorlunda bra. Modellen har $R^2 = 0.179$, $R^2_{adj} = 0.153$ och $p - värde = 1.176 \cdot 10^{-7}$. Modellen där principalkomponenterna skapats utan dummy variablerna är således bättre än modellen där dummy variablerna var delaktiga i skapandet av principalkomponenterna. Det visar sig även att stegvis variabelselektion resulterar i samma variabler oavsett om man använder 2, 4 eller 5 principalkomponenter som förklarande variabler i den multipla linjära regressionsmodellen. Slutligen ska även prediktionsförmågan för modell (34) testas med *LOOCV*. Resultatet av detta syns i tabellen nedan.

Tabell 4: Prediktionsförmåga med LOOCV för modell (34)

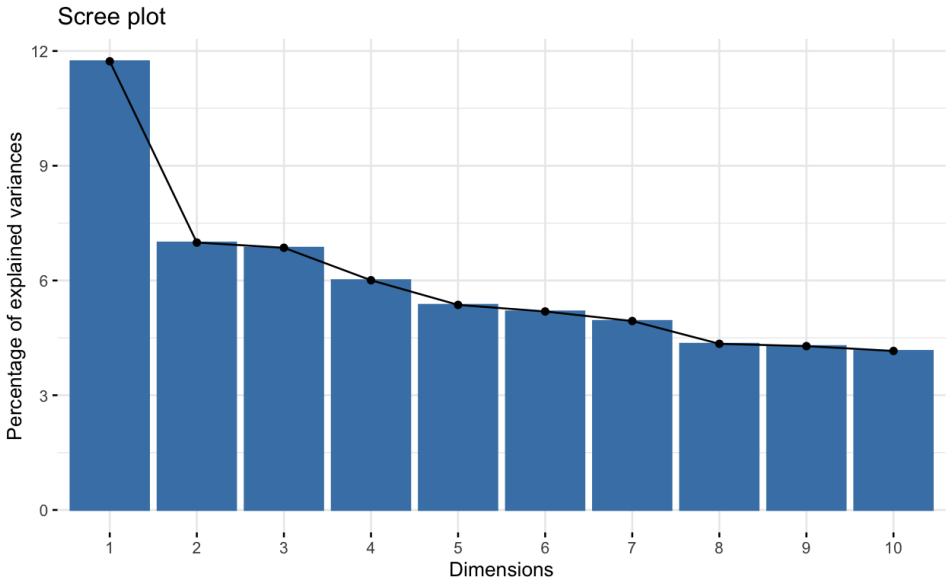
	R^2	RMSEP	MAEP
Modell (34)	0.113	0.374	0.286

Tabell 4 visar att prediktionsförmågan för modell (34) är likvärdig med prediktionsförmågan för modell (32). Anpassningsmåtten är dock bättre för modell (34) än för modell (32) så vi konstaterar att modell (34) är den bästa erhållna *PCR* modellen. Innan vi går vidare till modelleringen av *GAM* ska vi testa ytterligare ett sätt att hantera dummy variablerna på.

4.2.3 FAMD

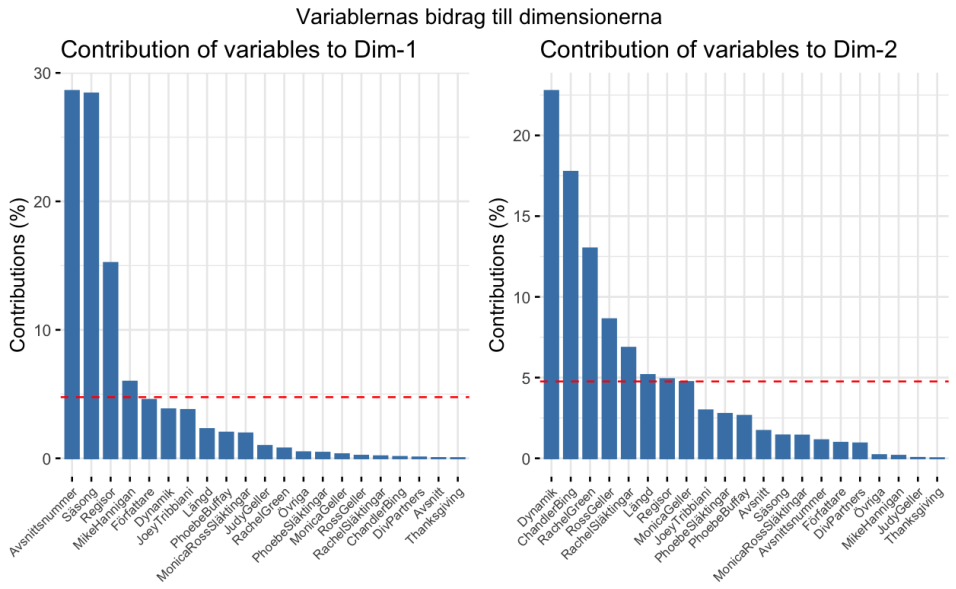
$FAMD$ algoritmen ska appliceras på numeriska samt kategoriska variabler. Datamaterialet som behandlats hittils innehåller 9 dummy variabler som representerar de 4 kategoriska variablerna **Regissör**, **Manusförfattare**, **Dynamik** samt **Thanksgiving**. I denna modell återställer vi dummy variablerna så att de återigen representeras av de fyra kategoriska variablerna. Vi applicerar sedan $FAMD$ på datamaterialet som nu innehåller 21 variabler.

Likt tidigare väljs antalet principalkomponenter efter armbågspunkten i PVE plotten. PVE plotten syns nedan, observera här att R:s inbyggda $FAMD$ funktioner talar om dimensioner istället för principalkomponenter.



Figur 11: PVE plott

Armbågspunkten i plotten i figur 11 sker vid dimension 2. Vi kommer således ställa upp en modell med de två första dimensionerna. Plottarna nedan visar de förklarande variablernas bidrag till respektive dimension.



Figur 12: Variablernas bidrag till dimensionerna

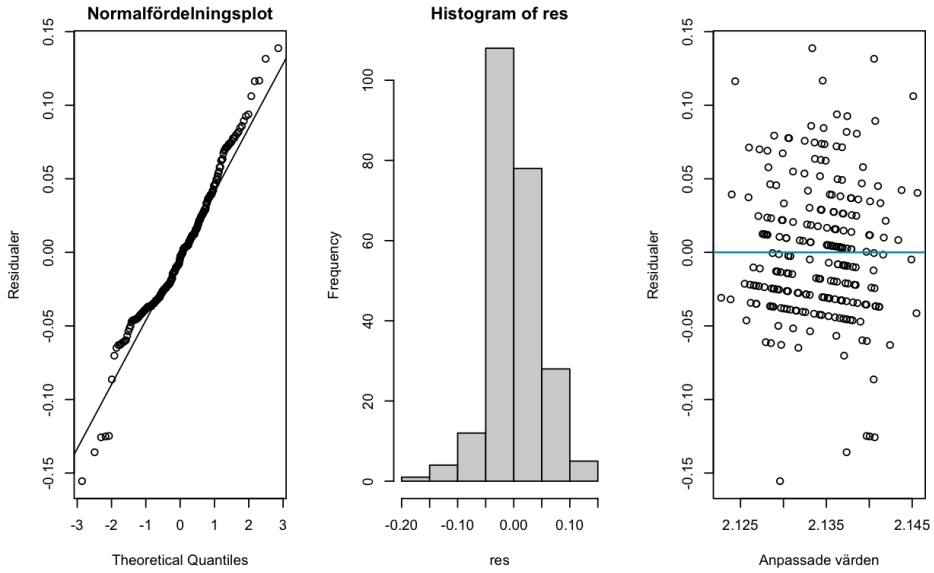
Låt oss nu ställa upp en multipel linjär regressionsmodell med de två första dimensionerna. Modellen kan skrivas:

$$IMDb_i = \alpha + \beta_1 \cdot Dim1_i + \beta_2 \cdot Dim2_i + \epsilon_i. \tag{35}$$

Modell (35) har $R^2 = 0.011$, $R^2_{adj} = 0.002$ samt $p - värde = 0.280$ och residualer som inte uppfyller kraven för feltermerna. Några olika transformationer testas där $log(IMDb)$ visar sig vara den bästa. En modell med responsvariabel $log(IMDb)$ skrivs:

$$log(IMDb_i) = \alpha + \beta_1 \cdot Dim1_i + \beta_2 \cdot Dim2_i + \epsilon_i. \tag{36}$$

Residualplottarna för modell (36) ges nedan i figur 13.



Figur 13: Residualplottar för modell (36)

Modell (36) har $R^2 = 0.010$, $R^2_{adj} = 0.002$ och $p - värde = 0.299$. Prediktionsförmågan för modell (36) ges i tabellen nedan:

Tabell 5: Prediktionsförmåga med LOOCV för modell (36)

	R^2	RMSEP	MAEP
Modell (36) med $log(IMDb)$	0.002	0.047	0.037
Modell (36) med IMDb	0.002	0.399	0.310

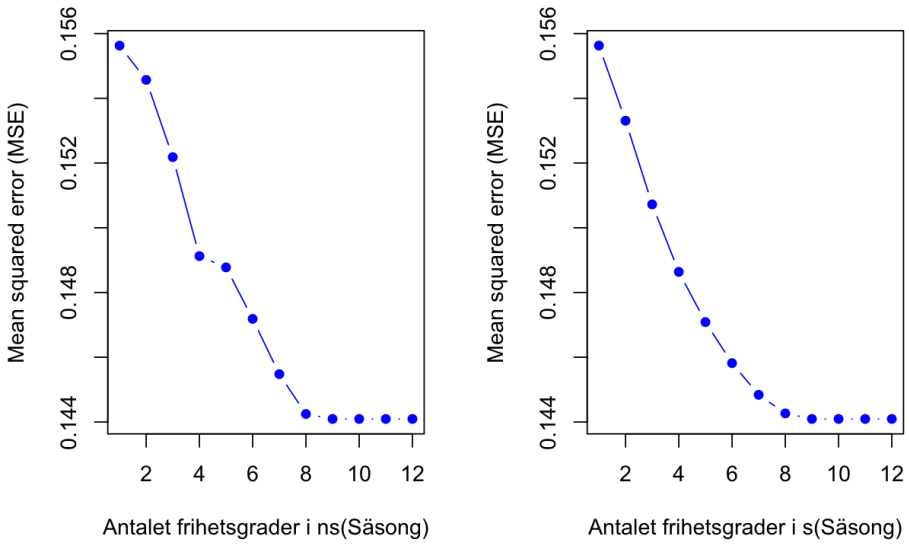
Vi har nu testad alla modeller som antar att det finns ett linjärt samband mellan de förklarande variablerna och responsvariabeln. Det har således blivit dags att testa en modell där vi inte antar att det finns ett linjärt samband. I nästa avsnitt testas en generaliserad additiv modell där icke linjära funktioner är tillåtna på variablerna.

4.3 Generaliserad additiv modell

Den sista modellen som ska testas är en generaliserad additiv modell där icke linjära funktioner är tillåtna på variablerna. I detta arbete kommer naturliga samt *smooth splines* testas på de numeriska variablerna och steg funktioner testas på de kategoriska variablerna i den generaliserade additiva modellen. Innan modellen kan ställas upp måste antalet frihetsgrader för de naturliga splinsen och de effektiva frihetsgraderna för *smooth splinsen* bestämmas.

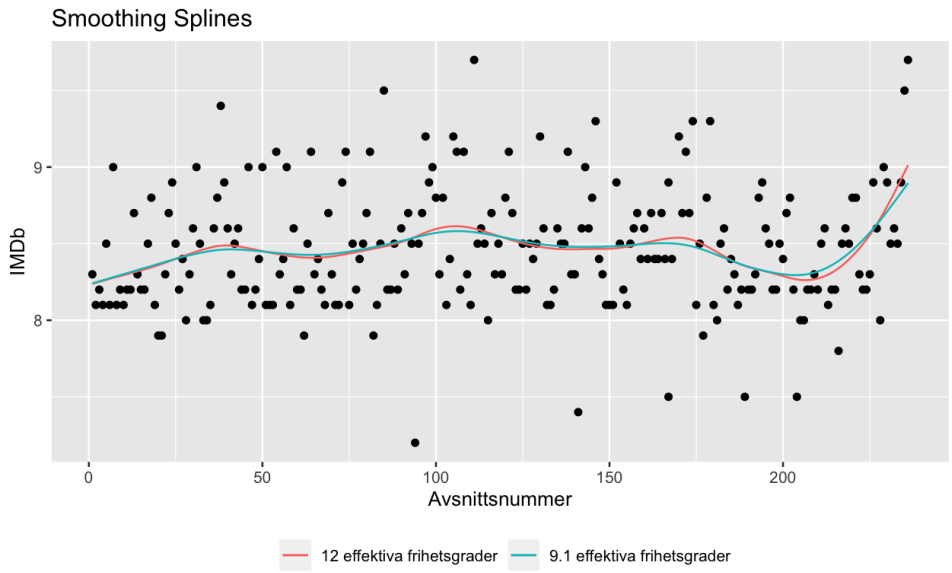
4.3.1 Antalet (effektiva) frihetsgrader

Antalet (effektiva) frihetsgrader som ska användas för respektive variabel avgörs med LOOCV. I korsvalideringen väljer man antalet (effektiva) frihetsgrader som ger lägst $MSEP$ ($Mean\ Squared\ Error\ of\ Prediktion$). För respektive variabel skapas plottar likt dem nedan i figur 14.



Figur 14: MSE för olika antal frihetsgrader i naturliga (ns) och effektiva frihetsgrader i smooth splines (s) på Säsong

Figur 14 visar att 9 (effektiva) frihetsgrader ger lägst MSE för naturliga och smooth splines på säsong. Nio stycken effektiva frihetsgrader här ger $\lambda = 0.0004$. Man kan även låta programmet välja nivån av *smoothness* genom *LOOCV* och sedan beräkna motsvarande λ och antal effektiva frihetsgrader. I *smooth* splines på variablerna har vi således två olika val av effektiva frihetsgrader. Plottar, likt dem nedan, skapas för respektive variabel för att avgöra antalet effektiva frihetsgrader som ska användas i *smooth* splinsen på variabeln.



Figur 15: Smooth splines på avsnittsnummer med 12 och 9.1 effekitva frihetsgrader

Vi erhåll att 12 effektiva frihetsgrader gav lägst MSE och 9.1 effektiva frihetsgrader var antalet frihetsgrader som motsvarade programmets valda *smoothness* level. I figur 15 ser man att de olika kurvorna inte skiljer sig något nämnvärt och eftersom 9.1 effektiva frihetsgrader ger en enklare modell väljer vi 9.1 effektiva frihetsgrader för smooth spline på avsnittsnummer. Samma steg och resonemang görs för alla variablerna och resultatet av detta syns i tabellen nedan.

Tabell 6: Antalet (effektiva) frihetsgrader för respektive variabel

Variabel	Antalet frihetsgrader i naturliga splines	Antalet effektiva frihetsgrader i smooth splines
Avsnittsnummer	11	9.1
Säsong	9	9
Avsnitt	12	12
Längd	2	2
ChandlerBing	6	10.3
JoeyTribbiani	10	12
MonicaGeller	10	12
PhoebeBuffay	12	12
RachelGreen	12	12
RossGeller	12	12
MonicaRossSläktingar	6	7
PhoebeSläktingar	3	3
RachelSläktingar	1	3
JudyGeller	2	4
MikeHannigan	1	5
DivPartners	11	10
Övriga	11	8

Nu när antalet (effektiva) frihetsgrader är bestämt kan man ställa upp den generaliserade additiva modellen.

4.3.2 Modellen

Vi ska nu ställa upp en generaliserad additiv modell. Vi genomför då en serie ANOVA test för att avgöra hur respektive variabel ska behandlas i modellen. För respektive numerisk variabel testas följande modeller:

- **M1:** modell utan variabeln,
- **M2:** modell med linjär funktion på variabeln,
- **M3:** modell med naturlig spline på variabeln,
- **M4:** modell med smooth spline på variabeln.

För respektive dummy variabel testas följande modellen:

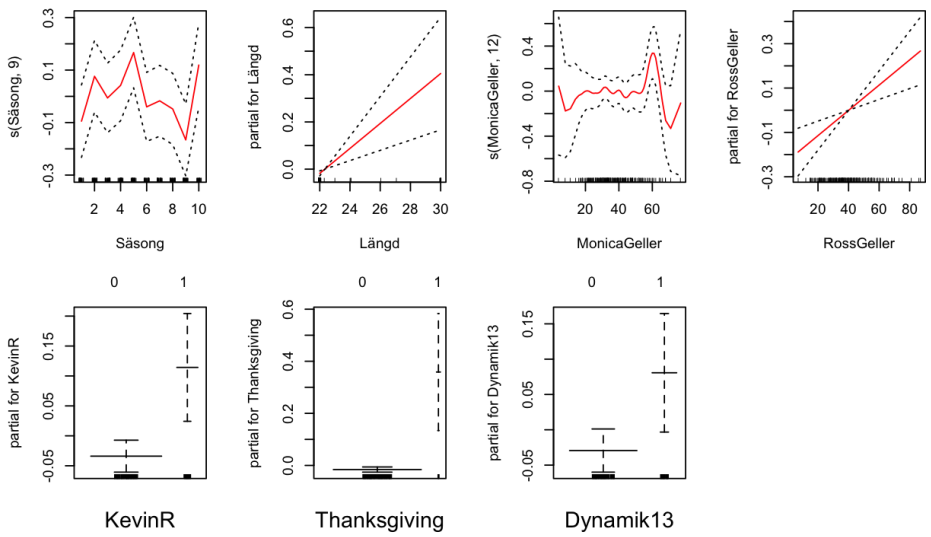
- **M1:** modell utan variabeln,
- **M2:** modell med steg funktion på variabeln.

I ANOVA testen testas nollhypotesen att modell $M1$ kan förklara datamaterialet mot mothypotesen att en mer komplex modell $M2$ behövs. I modellerna $M3$ och $M4$ ovan används antalet (effektiva) frihetsgrader enligt tabell 6. Resultatet av ANOVA testen är följande modell:

$$IMDb = \alpha + f_1(Säsong) + f_2(Längd) + f_3(MonicaGeller) \tag{37}$$

$$f_4(RossGeller) + f_5(KevinR) + f_6(Thanksgiving) + f_7(Dynamik13) + \epsilon$$

där f_j är smooth splines på Säsong samt MonicaGeller, linjära funktioner på Längd samt RossGeller och steg funktioner på dummy variablerna. Plottar för funktionerna i modell (37) syns nedan.



Figur 16: Plottar över de anpassade funktionerna och standardfelen i modell (37)

Plottarna i figur 16 visar att, om resterande variabler hålls fixerade, så resulterar de första och sista säsongerna i högre IMDb-betyg. IMDb-betyget ökar med avsnittslängden och med antalet repliker Ross har. IMDb-betyget verkar vara relativt konstant med antalet repliker som Monica har men det varierar kraftigt vid få och många antal repliker. IMDb-betyget ökar även för avsnitt som regisserats av Kevin, för special Thanksgiving avsnitt och för avsnitt som är centrerade runt en historia om Chandler och Monica.

Till sist ska predikteringsförmågan för modell (37) utvärderas med LOOCV. Resultatet av detta syns nedan i tabell 7.

Tabell 7: Prediktionsförmåga med LOOCV för modell (37)

	R^2	RMSEP	MAEP
Modell (37)	0.114	0.384	0.300

Alla modeller har nu modellerats så det är dags att jämföra dem, detta görs i nästa avsnitt.

5 Resultat

Det är nu dags att jämföra modellerna. De utvalda modellerna som ska jämföras är:

- **MLR** (30)**:** Multipel linjär regressionsmodell med Intercept, KevinR, MichaelR, AlexaM, Längd, Thanksgiving, Dynamik56, Dynamik13 och RossGeller,
- **PCR** (34)**:** Principalkomponent regression med PC2, KevinR, AlexaM, Thanksgiving, Dynamik56, Dynamik12 och Dynamik13,
- **FAMD** (36)**:** Faktor analys av blandad data med två dimensioner i den multipla linjära modellen,
- **GAM** (37)**:** Generaliserad additiv model med *smooth splines* på Säsong samt Monica Geller, linjära funktioner på Längd samt RossGeller och steg funktioner på KevinR, Thanksgiving samt Dynamik13.

Mått på prediktionsförmågan med LOOCV för de olika modellerna syns nedan i tabell 8.

Tabell 8: Prediktionsförmåga med LOOCV för modellerna

	R^2	RMSEP	MAEP
MLR (30)	0.147	0.367	0.289
PCR (34)	0.113	0.374	0.286
FAMD (36)	0.002	0.399	0.310
GAM (37)	0.114	0.384	0.300

Prediktionsförmågorna för de olika modellerna är relativt likvärdiga även fast den multipla linjära regressionsmodellen ger bäst prediktioner och FAMD ger sämst. Kriterierna för feltermerna i den multipla linjära regressionsmodellerna var dock inte helt uppfyllda och modellen innehåller variabler som inte är signifikanta på 5% signifikansnivå. Kriterierna för feltermerna i principalkomponent regressionsmodellen var inte heller helt uppfyllda. Vi drar

därmed slutsatsen att GAM är den mest lämpliga modellen för detta datamaterial och ändamål. I tabellen nedan ges parameterskattningarna för modellerna MLR och PCR.

Tabell 9: Parameterskattningar för modellerna MLR och PCR

Variabel	MLR (30)	PCR (34)
Intercept	7.0820	8.3026
Avsnittsnummer	-	0.0017
Säsong	-	0.0016
Avsnitt	-	0.0013
Längd	0.0473	0.0051
GaryR	-	-
KevinR	0.1612	0.1842
MichaelR	0.1232	-
AlexaM	0.1688	0.1688
AndrewTedM	-	-
Thanksgiving	0.3982	0.3350
Dynamik56	0.0848	0.1023
Dynamik12	-	0.1310
Dynamik13	0.1293	0.1564
ChandlerBing	-	0.0015
JoeyTribbiani	-	0.0027
MonicaGeller	-	0.0087
PhoebeBuffay	-	-0.0111
RachelGreen	-	0.0147
RossGeller	0.0046	0.0156
MonicaRossSläktingar	-	0.0092
PhoebeSläktingar	-	-0.0152
RachelSläktingar	-	0.0030
JudyGeller	-	0.0095
MikeHannigan	-	-0.0053
DivPartners	-	-0.0024
Övriga	-	-0.00444

De estimerade parameterskattningarna för den generaliserade additiva modellen kan tolkas ur figur 16. Tabell 9 och figur 16 visar att:

- IMDb-betyget ökar med säsongen, avsnittet inom säsongen samt totala mängden avsnitt. Enligt GAM sjunker IMDb-betyget kraftigt mellan säsong 5 och 9 men i allmänhet går det ändå uppåt,
- IMDb-betyget ökar med avsnittslängden vilket stöds av alla modellerna,

- IMDb-betyget ökar för avsnitt som regisserats av Kevin Bright vilket även det stöds av alla modellerna. IMDb-betyget ökar även för avsnitt som regisserats av Michael Lembeck eller skrivits av Alexa Junge,
- IMDb-betyget ökar för Thanksgiving avsnitt vilket stöds av alla modellerna,
- IMDb-betyget ökar för avsnitt som följer en historia om Rachel och Ross, Chandler och Joey samt Chandler och Monica. Här visar det sig att dynamiken Chandler och Monica associeras med högre ökning av IMDb-betyget än de andra dynamikerna,
- IMDb ökar även med antalet repliker som majoriteten av karaktärerna säger. Det sjunker dock ju mer repliker Phoebe, Phoebes släktingar, Mike, diverse partners och övriga karaktärer säger. IMDb-betyget varierar även mycket när Monica har få eller många repliker men i allmänhet går IMDb-betyget ned ju fler repliker Monica har.

6 Diskussion

6.1 Datamaterialet

I förberedelsen av datamaterialet gjordes en del ändringar i de urspungliga datamaterialen. De kategoriska variablerna i de urspungliga datamaterialen omvandlades till numeriska eller dummy variabler där respektive unik observation i den kategoriska variabeln fick en egen variabel. Gränserna för vilka variabler som sedan inkluderades i det slutgiltiga datamaterialet valdes något godtyckligt. Ett annat val av gränser i denna förberedelse hade kunnat påverka modellerna och resultatet.

Man hade även kunnat undersöka en annan responsvariabel, exempeltvis antalet tittare i Amerika som fanns tillgänglig i ett av de förberedande datamaterialen. I detta arbete ville vi dock fokusera på en responsvariabel och då valdes IMDb-betyget istället för antalet tittare. Anledningen till att IMDb-betyget valdes istället för antalet tittare är för att arbetet går ut på att beskriva vad som påverkar hur bra ett Friends-avsnitt anses vara och IMDb-betyget är just ett mått på detta. Antalet tittare har rimligtvis en viss korrelation med hur bra ett avsnitt anses vara men den korrelationen tros vara lägre än korrelationen mellan hur bra ett avsnitt anses vara och dess IMDb-betyget.

6.2 Multiple linjär regression

I den multipla linjära regressionsmodellen beslutades det att exkludera **Avsnittsnummmer** samt inkludera **Säsong** och **Avsnitt**. Detta val baserades på att säsongen och avsnittets ordning under serien var starkt parvist

korrelerade så en av dessa variabler behövde exkluderas. Resultatet blev likvärdigt oavsett vilken av dessa två variabler som exkluderades och då togs beslutet av inkludera säsongen eftersom datamaterialet handlar om en serie. Man hade självklart kunnat välja att exkludera säsongen istället men detta hade som sagt gett likvärdigt resultat. Det hade dock eventuellt varit av intresse att exkludera alla tre variabler och inkludera en sammanslagning av dem. En sådan sammanslagning hade blivit mer komplicerad och svårtolkad men hade kunnat bidra med information som missades när **Avsnittsnummer** exkluderades.

I valet av kriterie att använda i cooks avstånd valdes kriteriet $3 \cdot mean(D_i)$ eftersom det gav modellen med bäst $RMSE$, R^2 och MAE. I detta arbete valde vi att inte exkludera de avvikande observationerna som identifierades av Cooks avstånd eftersom vi ansåg att de innehöll viktig information och för att det är svårt att jämföra modeller med olika observationer. Ifall man hade gjort om arbetet hade man kunnat använda något av de andra kriterierna som testades för Cooks avstånd alternativt testat hur de andra modellerna påverkas av att exkludera de observerade avvikande observationerna. Den multipla linjära regressionsmodellen blev bättre utan de avvikande observationerna, ifall man hade testat och upptäckt att detsamma gällde alla modeller hade det kunnat vara anledning nog för att exkludera dem ur datamaterialet.

6.3 Principal komponent regression

I PCR modellerna bestämde man antalet principalkomponenter baserat på armbågspunkten i PVE plottarna. Det visade sig, för respekitve PCR modell, att armbågspunkterna bland annat skedde vid principalkomponent 2 samt principalkomponent p där p är antalet skapade principalkomponenter från de p förklarande variablerna. I dessa modeller valdes 2 principalkomponenter. Bättre resultat hade eventuellt kunnat uppnås ifall man valde p principalkomponenter men då detta skulle resultera i en vanlig multipel linjär regressionsmodell och PCR är en dimensionsreduceringsteknik inkluderades detta inte i detta arbete.

Det har även nämnts flertalet gånger att PCR kan vara olämpligt att applicera på dummy variabler. Man kan således argumentera för att en sådan modell inte bör testas överhuvudtaget. PCR modellen med dummy variablerna inkluderades dock i detta arbete för att kunna jämföra olika modeller där dummy variablerna behandlats på olika vis.

Ifall man hade gjort om detta arbete hade det varit intressant att undersöka andra dimensionsreduceringstekniker. Man hade exemepeltvis kunnat testa $PLSR$ (*Partial Least Squares Regression*) vilket, likt PCR, lämpar

sig väl på datamaterial med många förklarande variabler och multikollinära variabler.

6.4 Generalized additive model

Fördelen med *GAM* är att man kan använda ett flertal olika funktioner på variablerna. I detta arbete användes steg funktioner, linjära funktioner, naturliga *splines* och *smooth splines*. Ifall man hade gjort om arbetet hade det varit intressant att undersöka fler icke linjära funktioner, man hade exempeltvis kunnat testa lokal regression.

Antalet frihetsgrader i de naturliga *splinsen* valdes med *LOOCV* där vi lät frihetsgraderna variera mellan 1 och 12 och sedan valde det antal som gav lägst MSE. Antalet effektiva frihetsgrader i *smooth* splinsen valdes genom att jämföra antalet effektiva frihetsgrader som hittades genom två olika metoder. Valet mellan dessa metoder var inte, i alla fall, självklart då den ena kurvan kunde vara lite krokig medan den andra inte ansågs anpassa datamaterialet bra. Det fanns således fall där valet stod mellan risken att överanpassa samt underanpassa datamaterialet. Ifall man hade gjort om arbetet hade det varit intressant att testa ytterligare två generaliserade additiva modeller med *smooth* splines. En modell med antalet effektiva frihetsgrader som valts genom *LOOCV* och en modell med effektiva frihetsgrader som motsvarar programmets valda *smoothness* nivå.

7 Appendix

7.1 Förberedelse av de fem datamaterialen

Tabell 10: Ett litet urval av datamaterialet med repliker

friends					
text	speaker	season	episode	scene	utterance
There's nothing to tell! He's just some guy I work with!	Monica Geller	1	1	1	1
C'mon, you're going out with the guy! There's gotta be something wrong with him!	Joey Tribbiani	1	1	1	2
All right Joey, be nice. So does he have a hump? A hump and a hairpiece?	Chandler Bing	1	1	1	3

Datamaterialet gjordes om så att respektive unik observation i variabeln **speaker** fick en egen variabel med observationer som beskriver antalet repliker i respektive avsnitt. Resultatet av detta var ett datamaterial med 702

variabler som, efter reducering, gav ett datamaterial med 15 variabler. Detta datamaterial reducerades genom att exkludera svårtolkade variabler och variabler med mindre än fem observationer samt slå samman variabler som hade mellan fem och tjugo observationer.

Tabell 11: Ett litet urval av datamaterialet med vännernas dynamik

friendsdata			
epseason	epnum	epname	dynamics
1	1	The One Where Monica Gets a Roommate	3
1	1	The One Where Monica Gets a Roommate	5
1	1	The One Where Monica Gets a Roommate	56

Variabeln **dynamics** i datamaterialet med vännernas dynamik betyder att avsnittet följer en historia om karaktären / karaktärerna siffran motsvarar. Här motsvarar 1 Chandler, 2 Joey, 3 Monica, 4 Phoebe, 5 Rachel och 6 Ross. Även i detta datamaterial fick respektive unik observation i variabeln **dynamics** en egen variabel. Resultatet av detta var ett datamaterial med 43 variabler som, efter reducering, gav ett datamaterial med 6 variabler. Detta datamaterial reducerades genom att exkludera svårtolkade variabler och variabler med mindre än 30 observationer.

Tabell 12: Ett litet urval av datamaterialet med manusförfattare och regissörer

friends-tv-episodes							
ep_num	sea	ep	title	directed_by	written_by	air_date	us_viewers
1	1	1	”The Pilot”	James Burrows	David Crane, Marta Kauffman	September 22, 1994	21.5
2	1	2	”The One with the Sonogram at the End”	James Burrows	David Crane, Marta Kauffman	September 22, 1994	20.2
3	1	3	”The One with the Thumb”	James Burrows	Jeffrey Astrof, Mike Sikowitz	October 6, 1994	19.5

Precis som tidigare fick respektive unik observation i **directed_by** och **written_by** sin egna variabel. Resultatet av detta var ett datamaterial med 135 variabler som, efter reducering, gav ett datamaterial med 11 variabler. Detta datamaterial reducerades genom att exkludera de manusförfattare som hade mindre än 10 observationer och de regissörer som hade mindre än 20 observationer.

Tabell 13: Ett litet urval av datamaterialet med allmänna avsnittsegenskaper

friends_episodes_v3								
Air_year	Sea	Ep	Title	Duration	Summary	Director	Stars	Votes
1994	1	1	The Pilot	22	Monica and the gang introduce Rachel to the "real world" after she leaves her fiancé at the altar.	James Burrows	8.3	7440
1994	1	2	The One with the Sonogram at the End	22	Ross finds out his ex-wife is pregnant. Rachel returns her engagement ring to Barry. Monica becomes stressed when her and Ross's parents come to visit.	James Burrows	8.1	4888
1994	1	3	The One with the Thumb	22	Monica becomes irritated when everyone likes her new boyfriend more than she does. Chandler resumes his smoking habit. Phoebe is given 7000 when she finds a thumb in a can of soda.	James Burrows	8.2	4605

Tabell 14: Ett litet urval av datamaterialet om Thanksgiving avsnitt

Season	Episode	Thanksgiving
1	1	0
1	2	0
1	3	0

7.2 Tabeller med små urval av det slutgiltiga datamaterialet

Tabell 15: Ett litet urval av de 9 första förklarande variablerna

Avsnittsnummer	Säsong	Avsnitt	Längd	GaryR	KevinR	MichaelR	AlexaM	AndrewTedM
1	1	1	22	0	0	0	0	0
2	1	2	22	0	0	0	0	0
3	1	3	22	0	0	0	0	0

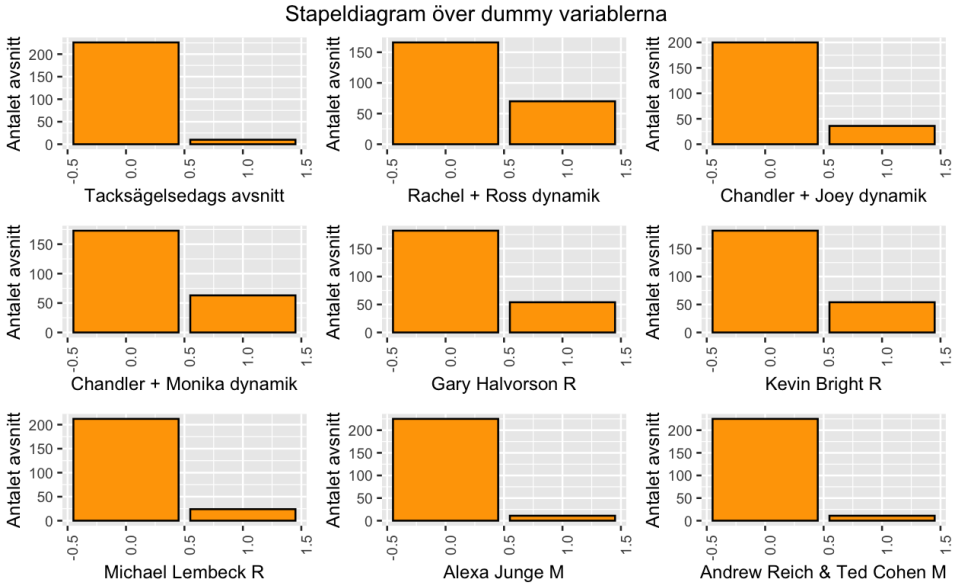
Tabell 16: Ett litet urval av de nästa 9 förklarande variablerna

Thanksgiving	Dynamik56	Dynamik12	Dynamik13	Chandler Bing	Joey Tribbiani	Monica Green	Phoebe Buffay	Rachel Geller
0	1	0	0	48	46	77	24	50
0	1	0	0	18	10	29	15	38
0	0	0	0	51	49	64	52	45

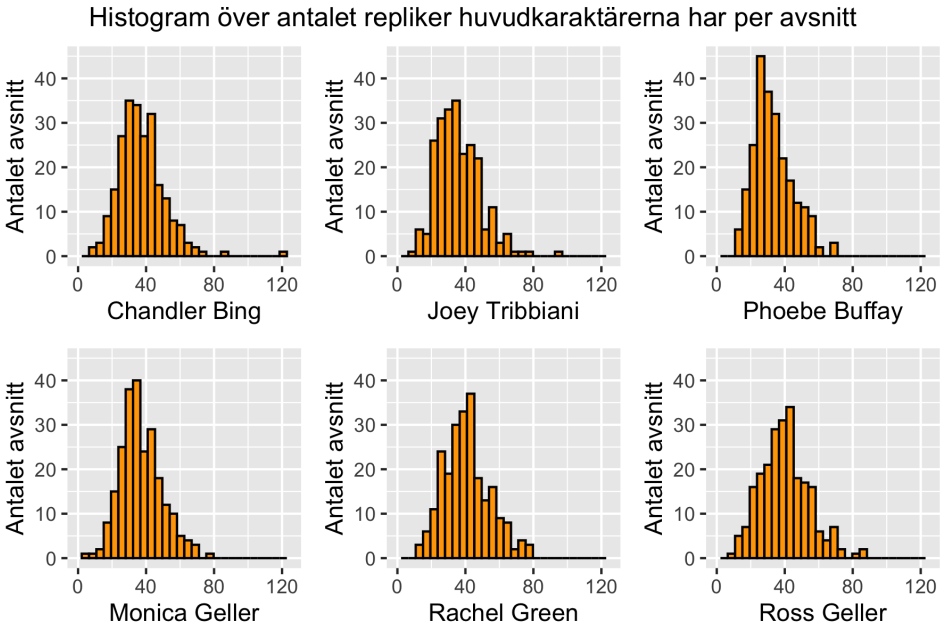
Tabell 17: Ett litet urval av de 8 resterande förklarande variablerna och responsvariabeln

Ross Geller	MonicaRoss Slätkingar	Phoebe Släktingar	Rachel Släktingar	Judy Geller	Mike Hannigan	Div Partners	Övriga	IMDb
50	0	0	0	0	0	0	0	8.3
69	30	0	0	10	0	16	0	8.1
52	0	0	0	0	0	0	0	8.2

7.3 Stapeldiagram och histogram över de förklarande variablerna

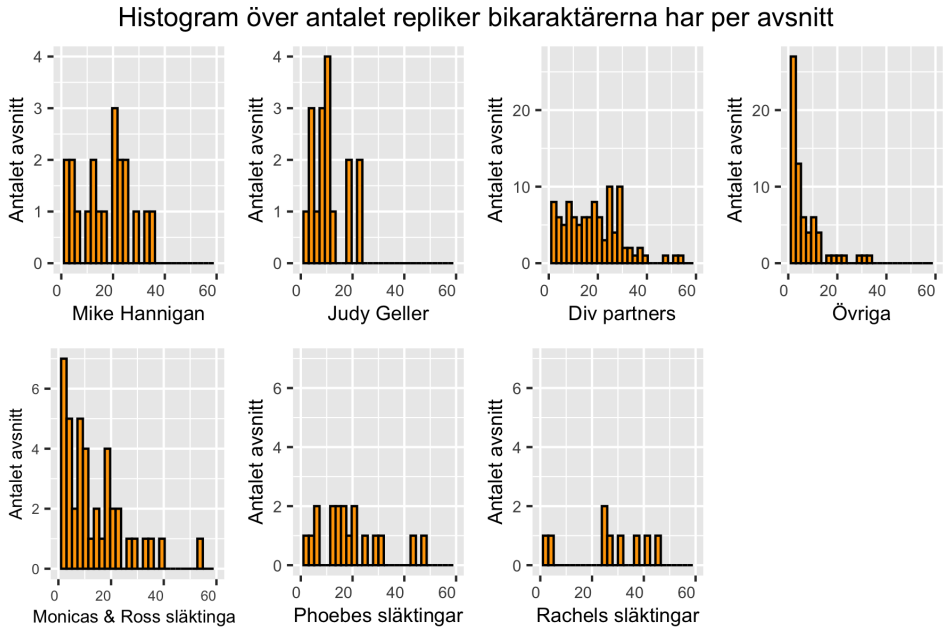


Figur 17: Stapeldiagram över datamaterialets dummy variabler



Figur 18: Histogram över huvudkaraktärernas repliker

Figur 18 visar att huvudkaraktärerna främst säger mellan 30 och 50 repliker per avsnitt. Figuren visar även att Chandler har störst spridning på antalet repliker han säger per avsnitt medan Phoebe har minst spridning.



Figur 19: Histogram över bikaraktärernas repliker där observationer som svarar mot 0 exkluderats

Figur 19 visar att Monicas och Ross släktingar har betydligt mycket fler repliker i serien än Phoebes och Rachels släktingar har. Figuren visar även att diverse partners till huvudkaraktärerna har en stor spridning på antalet repliker de säger i respektive avsnitt medan övriga bikaraktärer har en relativt låg spridning. Mike och Judy säger även ungefär lika många repliker under serien men Mike säger i genomsnitt mer repliker per avsnitt än Judy.

8 Referenser

Referenser

[1] Rolf Sundberg, *Lineära Statistiska Modeller*, augusti 2023.

[2] Sujay Kapadnis, "Friends Sitcom Dataset", www.kaggle.com, länk.

[3] Alex Albright, "Friends", https://github.com, länk.

[4] "List of all Friends TV episodes", https://dedolist.com, augusti 2020, länk

[5] Mohammad Reza Ghari och Moulik Dhade, "Friends Series Dataset", https://www.kaggle.com, länk.

[6] Josiah Soto, “Here’s Every Thanksgiving Episode of ’Friends’ to Watch This November”, www.thepioneerwoman.com, november 2023, länk

[7] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction To Statistical Learning*, Springer, 2013.

[8] Ping Yu, Lecture 05. Principal Components Analysis, The University of Hong Kong, länk

[9] William Blaufuks, “FAMD: How to generalize PCA to categorical and numerical data”, https://towardsdatascience.com, maj 2021, länk