Stockholms universitet

Modellering av antalet ''likes'' på Instagram med generaliserade linjära modeller

Josefin Andersson Senko

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:4
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2015:4**
http://www.math.su.se/matstat

Modellering av antalet "likes" på Instagram med generaliserade linjära modeller

Josefin Andersson Senko*

Juni 2015

Sammanfattning

Arbetet har till avsikt att svara på vad som påverkar antalet likes för en bild på Instagram. Målet är att hitta en modell som förklarar antalet likes på bästa sätt. För att hitta detta har jag samlat data från 45 olika kända personer. Jag har utgått från olika variabler varav fem är av dessa är kategoriska för att beskriva vilken typ av bild som avses. För att finna olika modeller används metoder i multipel linjär regression samt generaliserade linjära modeller. För att avgöra vilken modell som beskriver data bäst samt för att reducera variabler används AIC samt PIT. Slutmodellen består av en GLM med negativ binomialfördelning. I denna modell ser vi att en bild på endast användaren själv ger flest antal likes. Bilder utan personer genererar minst antal likes. Taggar inverkar negativt på antalet likes.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: josefin.andersson.senko@gmail.com. Handledare: Tom Britton, Martin Sköld.

Modellering av antalet ”likes” på Instagram med generaliserade linjära modeller

Josefin Andersson Senko

10 maj 2015

Sammanfattning

Arbetet har till avsikt att svara på vad som påverkar antalet likes för en bild på Instagram. Målet är att hitta en modell som förklarar antalet likes på bästa sätt. För att hitta detta har jag samlat data från 45 olika kända personer. Jag har utgått från olika variabler varav fem är av dessa är kategoriska för att beskriva vilken typ av bild som avses. För att finna olika modeller används metoder i multipel linjär regression samt generaliserade linjära modeller. För att avgöra vilken modell som beskriver data bäst samt för att reducera variabler används AIC samt PIT. Slutmodellen består av en GLM med negativ binomialfördelning. I denna modell ser vi att en bild på endast användaren själv ger flest antal likes. Bilder utan personer genererar minst antal likes. Taggar inverkar negativt på antalet likes.

Abstract

The intention of this study is to analyse how different variables affect the number of likes on a picture on Instagram. The goal is to find a model that explains the number of likes in the best way. To find this I have used a number of different variables, five of these are categorical and are used to describe the image. In order to find the best suited model we used methods in multiple linear regression and generalized linear models. To determine which model that suits data best and reduce variables AIC and PIT values have been used. The final model is a generalized linear model with negativ binomial distrubution. In this model, we see that a picture of the user only gives the highest number of likes. Images without people generates the least number of likes. Tags have a negative impact on the number of likes.

Förord

Detta arbete är en uppsats i matematisk statistik. Uppsatsen motsvarar 15 hp och är skriven vid matematiska institutionen på Stockholms Universitet. Jag vill främst tacka mina handledare Tom Britton och Martin Sköld för hjälp, stöd och engagemang under arbetets gång. Jag vill tacka Rasmus Parkinson som bidragit med feedback. Jag vill även tacka Swedish Match för hjälp med tålamodet.

Innehållsförteckning

1 Introduktion 2
1.1 Bakgrund 2
1.2 Avgränsning 2
1.3 Mål och syfte 2
2 Data 3
2.1 Variabler 3
2.2 Insamling av data och avgränsningar 3
2.3 Mer om data 4
3 Teori & metod 6
3.1 Multipel linjär regression 6
3.1.1 Logaritmtransformation 6
3.1.2 Minsta kvadrat metoden 6
3.2 GLM – generaliserade linjära modeller 7
3.2.1 Slumpfaktor 7
3.2.2 Systematiskfaktor 7
3.2.3 Länkfunktion 7
3.3 Poissonfördelningen 8
3.3.1 Slumpfaktor 8
3.3.2 Länkfunktion 8
3.4 Negativ binomialfördelningen 8
3.4.1 Slumpfaktor 9
3.4.2 Länkfunktion 9
3.5 Normalfördelning 9
3.5.1 Slumpfaktor 9
3.5.2 Länkfunktion 9
3.6 Likelihood 10
3.7 Akaike information criterion 10
3.8 Probability integral transform 11
4 Resultat 11
4.1 AIC 11
4.2 Residualer och PIT 12
4.3 Skattningar 13
4.4 Sammanfattning av resultat och slutgiltig modell 14
5 Diskussion 15
5.1 Resultat och skattningar 15
5.2 Framtida forskning och brister 16
Appendix 18
Referenser 22

1 Introduktion

1.1 Bakgrund

Instagram är ett socialt media avsedd för mobiltelefoner som tillåter användaren att publicera bilder och korta videoklipp. Vid en uppladdning tillåts en bildtext som även kan innehålla olika klickbara taggar, dessa kan relatera till andra användare (texttag) eller till vanliga ord, vilket skapar en sökbarhet (hashtag). I bilden tillåts även bildtaggar som gör att användaren kan visa vilka användare som visas på bilden. Dessa bilder och videoklipp kan sedan kommenteras och gillas av andra användare. Användaren väljer själv personer att följa och ser endast dessa personers uppdateringar i sitt flöde, med senast publicerade bild först.

Ur statistiskt perspektiv finns fler studier på ämnet sociala medier, oftast mycket nishade. För att nämna några har studie på vilka nyckelord som används av nyhetskällor som genererar flest likes på Facebook av Wolf (2013) gjorts eller en studie där Rospigliosi & Greener (2014) gjort regression på hur man kan använda sociala medier i högre utbildningar. Bondarouk & Olivas-Luján (2013) studerar hur man kan använda sociala medier i management & marknads strategier. Bakhshi, Gilbert, Chang & Terveen (2013) har skrivit en artikel om vad som driver engagemanget på Pinterest. Wolf (2015) skriver om användning av hashtags på Instagrarm för att lokalisera och analysera publicerade bilder av julgranar och huruvida de är spam eller inte. Bakhshi, Shamma & Gilbert (2014) skriver om hur olika ansikten påverkar aktiviteten på Instagram.

Eftersom att sociala medier är ett ämne som hör till vår tid känns det aktuellt för mig titta närmare på. Bakhshi, Shamma & Gilbert (2014) kommer i sin artikel fram till att närvaron av ansikten engagerar oss på Instagram i avseende på kommentarer och likes av bilder. De kommer även fram till att olika typer av ansikten inte påverkar engagemanget. Det kan ändå tänkas att det finns en skillnad på vilket typ av ansikte som visas. Möjligtvis inte i avseende på kön och ålder. Men i vilken mån man känner till personen på bilden eller inte. Det kan tänkas att val av de personer man följer beror på de personer man känner personligen eller har ett intresse av. Därför kommer jag utveckla artikeln ovan genom att undersöka vilket typ av ansikte som är med. Detta i avseende om ägaren till kontot är med på bilden eller inte.

1.2 Avgränsning

Arbetet kommer till skillnad från artikeln av Bakhshi, Shamma & Gilbert (2014) inte behandla lika mycket data. Jag har valt att fokusera på endast kända personer, detta för att hitta personer som står inför liknande förutsättningar på sociala medier. Jag kommer även endast att utgå från antalet likes och inte både likes och kommentarer. Samtliga begränsningar argumenteras av tid och resursbrister.

1.3 Mål och syfte

Syftet med denna uppsats är att undersöka hur olika typer av bilder påverkar antalet likes. Målet med uppsatsen är att undersöka hur olika faktorer påverkar antalet likes för bilder på Instagram och hitta en modell som beskriver detta på bästa sätt. Jag har valt följande frågeställning att besvara

- Hur påverkas antalet likes av vilken typ av bild som visas?
- Vilken modell är den bästa modellen för att förklara antalet likes på en bild?

2 Data

2.1 Variabler

Som responsvariabel har $Y = antalet\ likes$ valts. För olika typer av bilder kommer jag använda faktornivåer representerade av dummyvariabler. Dessa antar värdet ett om bilden uppfyller önskad egenskap och noll om egenskapen saknas. För de faktornivåer som ska beskriva vilken typ av egenskap bilden uppfyller har jag valt följande dummys;

- x_1, Endast personen som äger kontot
- x_2, Personen som äger kontot och en person till
- x_3, Personen som äger kontot och flera personer
- x_4, Andra personer utan personen som äger kontot
- x_5, Inga ansikten.

x_1 och x_5 kan ses som extremer för dessa kategorier, vilket gör att x_3 ses som den neutrala variabeln. Därför kommer jag använda x_3 som basvariabel för dessa kategorier vid modelleringen. För att komplettera dessa variabler ovan har även följande variabler som är unika för varje bild eller person vid insamlingstillfället valts

- x_6, Hur många veckor en bild har legat ute
- x_7, Antalet hashtags
- x_8, Antalet taggar i bildtexten
- x_9, Antalet bildtaggar
- x_{10}, Antal följare för personen.

Även då alla personer som data samlats ifrån är kända och står för liknande förutsättningar, kan det ändå tänkas att personernas följare är olika mycket hängivna. Det vill säga, vissa personers följare har högre aktivitet än andra och att detta är trender som inte beskrivs av variablerna ovan. För att motverka att individuella variationer inverkar kommer vi att tilldela varje person en faktorvariabel. För att tydligt se vilken effekt varje användare har kommer vi använda den person som har lägst medelvärde för antalet likes som basvariabel.

Variabeln *antal följare* är högst relevant för antalet likes och kommer ses som en offset variabel. En variabel där parametern framför alltid är ett.

2.2 Insamling av data och avgränsningar

För att välja personer med samma förutsättningar har jag utgått från en lista med de 50 personer som hade flest följare i januari 2014 Roosevelt H.Q (2014). Det är då rimligt att anta att alla dessa personer är kända och står för liknande förutsättningar på sociala medier. Från dessa samlas sedan data från de 50 senast utlagda bilderna för varje konto.

Insamling av data har skett manuellt i februari 2015, data är unik för insamlingstillfället. Eftersom många kändisar har fans över hela världen har jag hoppat över det senaste dygnet hos varje användare. Detta för att inte få skeva observationer beroende på vad klockan är världen över. Då vi ignorerar de senaste 24 timmarna har hela världen fått chansen att se bilden. Förutom de senaste 24 timmarna har de senaste 50 upplagda bilderna bidragit till datamaterialet.

Tecknade bilder har ignorerats då dessa skiljer i kvalitet samt är svåra att placera de kategoriska variabler som har valts, vi bortser då från ca 20 bilder som står för mindre än 1% av materialet. Även videoklipp samt otydliga bilder där det är svårt att avgöra om personen är med eller inte har ignorerats. Vi tar bort ungefär lika många videoklipp som tecknade bilder, ett tjugotal. Gällande otydliga bilder har jag hoppat över ca 50 stycken totalt, ca 2 %.

Då insamling av data skedde valde jag att utesluta fem av de personer som presenteras i den lista jag utgått från. Detta av olika anledningar. Louis Tomlinson från One Direction har endast publicerat 24 inlägg totalt. Barack Obamas Instagram används mycket sällan och mest i perioder vid val, detta ger en skev bild av antalet likes då de publiceras flera bilder vid val och knappt en varannan månad övrig tid, detta ger Obama ett tidsspann på flera år och fler likes på de bilder som publiceras då det inte är valtider. Både Bruno Mars och Drake har för många bilder där det är svårt att avgöra om de själva är med på bilden eller inte, dessa två har uteslutits av argument för tidseffektivitet. Efter att ha reducerat Bruno Mars och Drake har vi endast ignorerat ett tjugotal otydliga bilder totalt. Talia Joy Castellano har avlidit och hennes Instagram används som en hyllning för henne, de bilder som publiceras är gamla bilder på henne. Det är oklart hur detta påverkar antalet likes, därför har även hon uteslutits från data. Efter dessa avgränsningar har vi alltså 45 olika personer att utgå från.

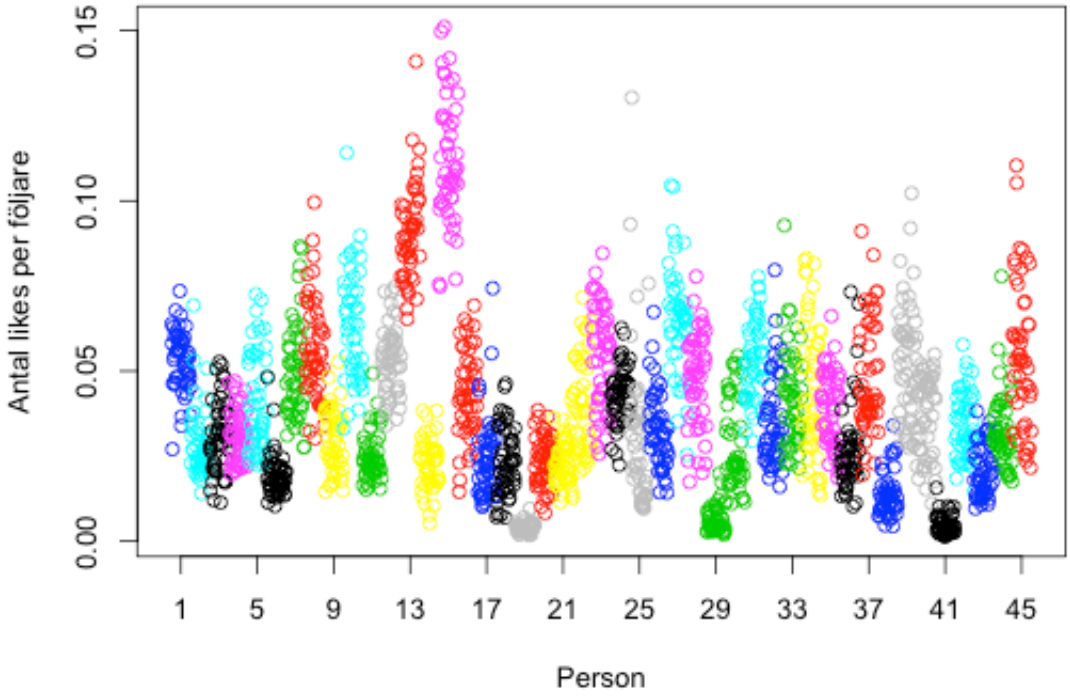
Eftersom att data att samlats in manuellt så har jag för varje person valt ut observation nummer 20 samt observation nummer 40, totalt 90 bilder från data. För att säkerställa korrekt inmatning av data har jag sedan låtit en testgrupp á två personer avgöra vilka egenskaper bilderna uppfyller. I Tabell 2.1 nedan ser vi att det finns tillräckligt många bilder i varje kategori för att försöka besvara frågeställningen om hur antalet likes påverkas av olika bilder.

Tabell 2.1. *Antal och andel bilder per bildkategori.*

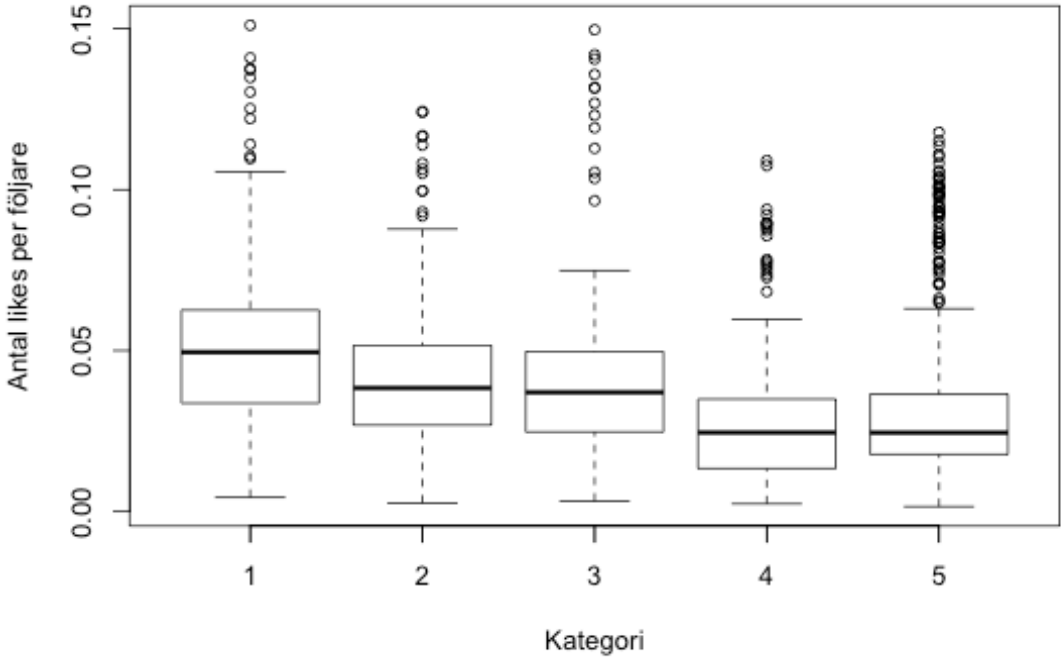
Typ av bild	Antal	Andel
Endast användaren	651	29 %
Användaren och en till person	416	18,5 %
Användaren och flera andra personer	282	12,5 %
Andra personer	226	10 %
Utan ansikten	675	30 %
Totalt	2250	100 %

2.3 Mer om data

Antalet likes skiljer sig mycket mellan personerna beroende på bland annat antalet följare, därför presenteras figurer nedan med antalet likes per följare. Figur 2.1 visar antal likes per följare i olika färger beroende på person. Index representerar vilken person observationerna avser, där ordningen på personerna redovisas i Tabell 1 i Appendix. Varje person har 50 observationer. Vi ser i figuren att individuella variationer verkar existera. Figur 2.2 visar en boxplot för de olika kategorierna. Även om skillnaderna är små ser det ut att finnas skillnader mellan de olika kategorierna. I Tabell 2.2 visas minimum, medel och maximum för olika variabler.



Figur 2.1. *Antal likes per följare där olika färger markerar observationerna från de olika personerna.*



Figur 2.2. *Boxplot för antal likes per följare för varje bildkategori.*

Tabell 2.2. *Min, medel och max värde för olika kategorier i vårt datamaterial.*

	Min	**Medel**	**Max**
Antal följare	1 900 000	10 437 777	25 000 000
Antal likes	3 555	428 811	2 090 367
Veckor på Instagram	0.14	6.01	103
Antal hashtags	0	0.70	10
Antal texttaggar	0	0.54	8
Antal bildtaggar	0	0.12	7

3 Teori & metod

Detta kapitel behandlar den teori och de modeller som behövs för analysen. Teorin för GLM kommer primärt från Agresti (2002, s 116-131) och Fox (2015, s 403-404), om inte annat anges är dessa de källor som används. För kapitlet Likelihood har Fox (2015, s 404) använts.

3.1 Multipel linjär regression

För responsvariabeln $Y = antalet\ likes$ kan vi använda multipel linjär regression för att hitta en modell som beskriver data. Detta även för att hitta skattningar för parametrarna $\alpha, \beta_1, \beta_2, \dots, \beta_9$ som beskriver vilken inverkan förklaringsvariablerna $x_{i1}, x_{i2}, \dots, x_{i9}$ har på antalet likes. Detta skrivs på formen

$$y_i = x_i\beta + \epsilon_i, i = 1,2, \dots, n$$

Vektorn $\beta = (\alpha, \beta_1, \beta_2, \dots, \beta_9)^T$ beskriver vilken effekt variablerna $x_i = (1, x_{i1}, x_{i2}, \dots, x_{i9})$ har på responsvariabeln y_i, där α är intercept. Index i beskriver vilken observation vektorerna x_i och y_i behandlar. Termern ϵ_i kallas för slumptermen och antas oftast vara oberoende, normalfördelade med väntevärde noll och konstant varians.

3.1.1 Logaritmtransformation

Om data inte passar en helt linjär modell kan vi göra en logaritmering av responsvariabeln. Detta bidrar även till att få residuerna mer jämnt spridda. Modellen kommer då istället bli

$$\log(y_i) = x_i\beta + \epsilon_i, i = 1,2, \dots, n.$$

Vi kommer att använda denna typ av multipel regression i denna studie då antalet likes varierar mycket i storlek. Multipel linjär regression behandlar kontinuerlig data, vilket egentligen inte stämmer överens med vårt datamaterial. Att notera är dock att en responsvariabel med värdet noll inte fungerar i regressionen ty $\log(0)$ är odefinierat. Detta är dock inget problem i vårt datamaterial då minsta antalet likes i datamaterialet är 3555. För läsare som inte är bekant med multipel linjär regression rekommenderas Andersson & Tyrcha (2014) för vidare läsning.

3.1.2 Minsta kvadrat metoden

Det statistiska programpaketet R använder minsta kvadrat metoden för att skatta parametrarna β i en multipel linjär regression. Med minsta kvadrat metoden skattas parametrarna genom att minimera $\sum_i (y_i - x_i\beta)^2$, alltså kvadraten på residualerna. Vi vill minimera avståndet mellan regressionslinjen och de observerade värdena. Residualerna kan skrivas som en ekvation på formen

$$e_i = y_i - x_i\beta.$$

Parametrarna β skattas genom ekvationen

$$\hat{\beta} = (x^T x)^{-1} xy.$$

Där x är en matris med rad vektorer $x_1, x_2, \dots, x_n$ och $y = (y_1, y_2, \dots, y_n)^T$. För vidare läsning och härledningen av minsta kvadrat metoden hänvisas läsaren till Andersson & Tyrcha (2014, s 17-35).

3.2 GLM – generaliserade linjära modeller

Generaliserade linjära modeller utökar vanliga regressionsmodeller till data som har andra fördelningar. Dessa består av tre olika komponenter; en slumpfaktor – det vill säga en sannolikhetsfördelning från exponentialfamiljen, en systematisk faktor och en länkfunktion. Dessa tre presenteras närmare nedan.

3.2.1 Slumpfaktor

Slumpfaktorn i en GLM består av en variabel Y med oberoende observationer från en fördelning i exponentialfamiljen. Exponentialfamiljen har följande form

$$f(y_i; \theta_i) = a(\theta_i) b(y_i) \exp\{y_i Q(\theta_i)\}.$$

Funktionen $Q(\theta_i)$ kallas den naturliga parametern. Inte alla fördelningar kan skrivas på denna form, de flesta vanliga fördelningar kan dock skrivas på den utökade formen av exponentialfamiljen, även kallad den exponentiella spridningsfamiljen. Dess fördelning skrivs

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}.$$

Vi kommer utgå från den utökade formen för att kunna skriva alla fördelningar på samma form. Vi kan jämföra parametrarna i den naturliga exponentialfamiljen med den som tillhör den utökade exponentialfamiljen, ϕ kallas för spridningsparametern.
$\frac{\theta_i}{a(\phi)}$ motsvaras av $Q(\theta_i)$, $\exp\left\{-\frac{b(\theta_i)}{a(\phi)}\right\}$ motsvaras av $a(\theta_i)$och $\exp\{c(y_i, \phi)\}$ motsvaras av $b(y_i)$.

3.2.2 Systematiskfaktor

Den systematiska faktorn i en GLM relaterar η_i till förklaringsvariablerna genom följande linjära funktion

$$\eta_i = x_i \beta, i = 1,2, \dots, n.$$

Där x_i är vektorn med förklaringsvariablerna för observation *i*.

3.2.3 Länkfunktion

För att relatera slumpfaktorn till den systematiska faktorn används en länkfunktion i GLM. Vi låter $E[Y_i] = \mu_i$, koppla ihop den systematiska faktorn och förklaringsvariablerna genom funktionen g, $\eta_i = g(\mu_i)$. Vi kallar g för länkfunktionen i GLM. Då vi har att $\eta_i = x_i \beta, \eta_i = g(\mu_i) = x_i \beta$ samt det faktum att g är inverterbar ger oss $\mu_i = g^{-1}(x_i \beta)$.

3.3 Poissonfördelningen

Den vanligaste och enklaste GLM modellen för räknedata, data som räknar antalet av något är Poisson. Poisson fördelningen har följande täthetsfunktion

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1,2 \ldots$$

där $\mu > 0$ och $E[Y] = Var(Y) = \mu$.

3.3.1 Slumpfaktor

Poisson fördelningen skrivas på den naturliga exponentialfamiljens form. Detta sker med $\theta = \mu, a(\mu) = \exp(-\mu), b(y) = \frac{1}{y!}$ och $Q(\mu) = \log(\mu)$. Översatt till den utökade formen kan vi skriva fördelningen på formen

$$f(y_i;\mu_i) = \exp\{y_i \log(\mu_i) - e^{\mu_i} + -\log(y_i!)\}$$

Detta uppfyller en utökad exponentialfamilj med $\theta = \mu, a(\phi) = 1, b(\theta) = e^{\mu}$ och $c(y,\phi) = -\log(y!)$.

3.3.2 Länkfunktion

Den naturliga parametern för Poisson är log, vilket ger länkfunktioen $\eta = \log(\mu)$, detta ger oss

$$\log(\mu) = x\beta.$$

Där β är en vektor med koefficienter för hur mycket varje variabel påverkar, x motsvarar vektorn innehållandes samtliga förklaringsvariabler. Här beskriver β inverkan av x på det logaritmerade värdet av y. Denna modell uppfyller

$$\mu = \exp(x\beta) = e^{x\beta}.$$

En ökning med en andel av något x_i ger en multiplikativ inverkan av e^{β_i} på μ_i. I vårt fall har vi antal följare som en offset variabel. Denna variabel alltid har ett konstant β med värdet ett. Vi får då en modell som uppfyller

$$\mu = \exp(x\beta + \log(antal\ följare)) = e^{x\beta} \cdot antal\ följare.$$

Vi modellerar alltså antalet likes per följare. Vid användning av log som länk istället för identiteten, alltså $\mu = \alpha + \beta x$ så får vi en ”rundare” regressionslinje istället för en linjär. När vi modellerar en variabel med Poisson så antar vi $E[Y] = Var(Y) = \mu$. Detta gäller dock inte alltid då vi ofta har överspridning i data vid antagandet av Poisson. Överspridning ger effekten av att p-värdena sänks och variablerna blir allt för lätt signifikanta. Överspridning uppstår alltså då $E[Y] < Var(Y)$. I dessa fall kan vi använda oss av en utökning av Poisson GLM, negativ binomialfördelning.

3.4 Negativ binomialfördelningen

Likt Poisson stöds den negativ binomialfördelningen endast av positiva tal, men mot för Poisson har negativ binomialfördelningen högre varians än väntevärde. Om detta är fallet är det bättre att anta att responsvariabeln är negativt binomialfördelad. Dess täthetsfunktion ges av

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y, y = 0,1,2, \ldots.$$

I denna funktion är k och μ parametrar. Γ är gamma funktionen som för heltalsvärda n definieras av $\Gamma(n+1) = n!$.Med negativ binomial fördelning får vi att $E[Y] = \mu$ samt variansen $Var(Y) = \mu + \frac{\mu^2}{k}$. Där k kallas spridningsparameter. När $\frac{1}{k} \to 0$ konvergerar fördelningen mot en Poisson fördelning. Spridningsparametern är okänd och måste skattas från data.

3.4.1 Slumpfaktor

Negativ binomialfördelningen uppfyller inte den naturliga exponentialfamiljens form. Efter logaritmering ser vi att vi kan skriva negativ binomialfördelningen på formen

$$f(y_i;k,\mu_i) = \exp\left\{y_i \log\left(1-\frac{k}{\mu_i+k}\right) + k \cdot \log\left(\frac{k}{\mu_i+k}\right) + \log\left(\frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)}\right)\right\}.$$

Detta uppfyller en utökad exponentialfamilj med $\theta = \mu, a(\phi) = 1$, $b(\theta) = k \cdot \log\left(\frac{k}{\mu+k}\right)$ och $c(y,\phi) = \log\left(\frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)}\right)$.

3.4.2 Länkfunktion

I R används log som standard länkfunktionen för negativ binomialfördelning. Detta innebär att vi får samma länkfunktion för negativ binomialfördelningen som vi fick för Poissonfördelningen ovan.

3.5 Normalfördelning

Det kan även tänkas att data passar en vanlig normalfördelning. Även då normalfördelningen behandlar kontinuerlig data vilket inte stämmer överens med datamaterialet. Dess fördelning ges av

$$f(y;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right).$$

3.5.1 Slumpfaktor

Normalfördelningen kan även denna skrivas som en vanlig exponentialfamilj för den intresserade läsaren skriver Liero & Silvelyn (2012, s 15) mer om detta. Efter omskrivningar får vi normalfördelningen på formen

$$f(y_i;\mu,\sigma) = \exp\left\{\frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}.$$

Normalfördelningen skrivs som utökad exponentialfamilj med följande motsvarigheter $\theta = \mu, \phi = \sigma^2, a(\phi) = \phi, b(\theta) = \frac{\theta^2}{2}$ och $c(y,\phi) = -\frac{1}{2}\left(\frac{y_i^2}{\phi} + \log(2\pi\phi)\right)$.

3.5.2 Länkfunktion

För att kunna jämföra de tre olika GLM modellerna så kommer vi använda log som länk även för normalfördelningen.

3.6 Likelihood

R använder likelihoodfunktionen L för att skatta parametrarna β i en GLM. Likelihoodfunktionen defineras som

$$L(\beta|y) = \prod_i P(Y_i = y_i).$$

Från denna får vi log likelihooden

$$l(\beta|y) = \log\big(L(\beta|y)\big) = \log\left(\prod_i P(Y_i = y_i)\right).$$

Eftersom att alla tre GLM modeller kan skrivas på den utökade exponentialfamiljens form så utgår vi från denna för att få en generell skattning som är giltig för de tre modellerna.

$$l(\beta|y) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i, \phi).$$

Lösningen för $\beta_1, \beta_2, \dots, \beta_k$ av log likelihoodfunktionen ges av

$$\frac{\partial l(\beta_j)}{\partial \beta_j} = \sum_i \frac{\partial l_i}{\partial \beta_j} = 0\,, j = 1,2, \dots, k.$$

För att hitta skattningar till parametrarna måste vi derivera likelihoodfunktionen med avseende på varje enskild parameter. Derivering av log likelihooden med kedjeregeln ger

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}, j = 1,2, \dots, k.$$

Vidare omskrivningar leder till

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i(\phi)v(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij}, j = 1,2, \dots, k$$

där $a_i(\phi)v(\mu_i)$ är variansen för modellen och $v(\mu_i) = b''(\theta_i)$. Detta ger oss maximum likelihoodskattningarna för parametrarna genom ekvationen

$$\sum_i \frac{y_i - \mu_i}{a(\phi)v(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij} = 0\,, j = 1,2, \dots, k.$$

För negativ binomial måste vi även skatta spridningsparametern. Denna skattas i R med hjälp av maximum likelihood. För läsaren som är intresserade av denna skattning hänvisas denne till Aragan, Eberly D. och Eberly S (1991).

3.7 Akaike information criterion

Då vi har flera modeller att jämföra använder vi AIC, *Akaike information criterion*. Vi definierar AIC som

$$AIC = -2(\log(L) - k).$$

Där k motsvarar antal parametrar i modellen och L maximum likelihooden för modellen. AIC belönar enklare modeller med färre parametrar och höga likelihood värden. AIC ses

som en mått på hur mycket information som har förlorats i modellen. Vi föredrar alltså den modellen med lägst AIC (Agresti, 2002, s 216).

Vi kan även använda AIC för att reducera parametrar. För att hitta den bäst anpassade modellen kommer vi använda funktionen stepAIC i R från paketet MASS av Venables och Ripley (2002). Funktionen utför en stegvis variabel selektion genom att testa vilken variabel som ger lägst AIC då vi reducerar den. I nästa steg väljer vi den modellen med variabeln borttagen och utför samma sak igen. StepAIC testar även att lägga till parametrar som har reducerats för att hitta den bäst anpassade modellen. Detta tills vi har hittat en modell där vi inte längre får lägre AIC vid reducering eller tillägg av variabler. Vi kommer inte kunna jämföra AIC mellan de modeller vi får i multipel regression och de vi får från GLM regressionerna.

3.8 Probability integral transform

För att undersöka huruvida de valda fördelningarna passar data eller ej kommer vi använda PIT, *probability integral transform* som Czado, Gneiting och Held presenterar (2009, s 1254-1256). För att få fram de transformerade värdena använder vi att Y har en av de antagna fördelningarna. Den kumulativa fördelningsfunktionen ges av $F_Y(y;\mu_i)$. Det gäller då att $Z_i = F_Y(y_i, \mu_i)$ kommer vara likformigt fördelat om den valda fördelningen är väl anpassad till data.

För att undersöka detta kommer vi plotta histogram för alla Z_i och undersöka huruvida de ser ut att följa en likformig fördelning. Då vi får ett U-format histogram visar detta att data har överspridning i den valda modellen eller att fördelningen är underspridd. Ett histogram som ser mer bergsliknande ut indikerar att data har underspridning alternativt att fördelningen har en överspridning.

4 Resultat

Med hjälp av R utvecklad av R Core Team (2015) har vi utfört regression med funktionen glm och glm.nb från paketet MASS utvecklat av Venables och Ripley (2002) för att undersöka vad som påverkar antalet likes för en bild på Instagram. Detta med modellen multipel linjär regression av de logaritmerade värdena av responsvariabeln Y samt de generaliserade linjära modeller som utgår från Poisson-, Negativ binomial- och Normalfördelningen.

4.1 AIC

I Tabell 5.1 presenteras de AIC som ges för de olika modellerna då vi använder samtliga variabler i regressionen. Att notera är dock att vi inte kan jämföra AIC för modeller som tillhör multipel linjär regression med regressioner inom generaliserad linjära modeller. Det vill säga modellerna kan endast jämföras med andra modeller inom samma regressionsmetod.

Tabell 5.1. *AIC för olika modeller med samtliga variabler.*

Modell	**AIC**
Logaritmerad respons	335
Poisson	47 564 359
Negativ Binomial	56 493
Normal	58 893

Efter att ha använt funktionen stepAIC i R på de tre modellerna preseterar R vilka variabler som bör tas bort för att minimerar AIC. För regressionen med den logaritmerade responsvariabeln tar vi bort $x_8 = antal\ bildtaggar$. För Poissonfördelningen får vi inget lägre AIC om vi tar bort någon variabel. För negativ binomialfördelningen tar vi bort $x_8 = antal\ bildtaggar$. För normalfördelningen tar vi bort $x_7 = antal\ hashtags$.

Vid justering av modellerna där vi reducerar de variabler som R föreslår får vi nya AIC-värden. Dessa presenteras i Tabell 5.2. Vi ser att de modeller som justerats får något bättre AIC. Vi kommer härifrån utgå från de nya justerade modellerna för varje fördelning.

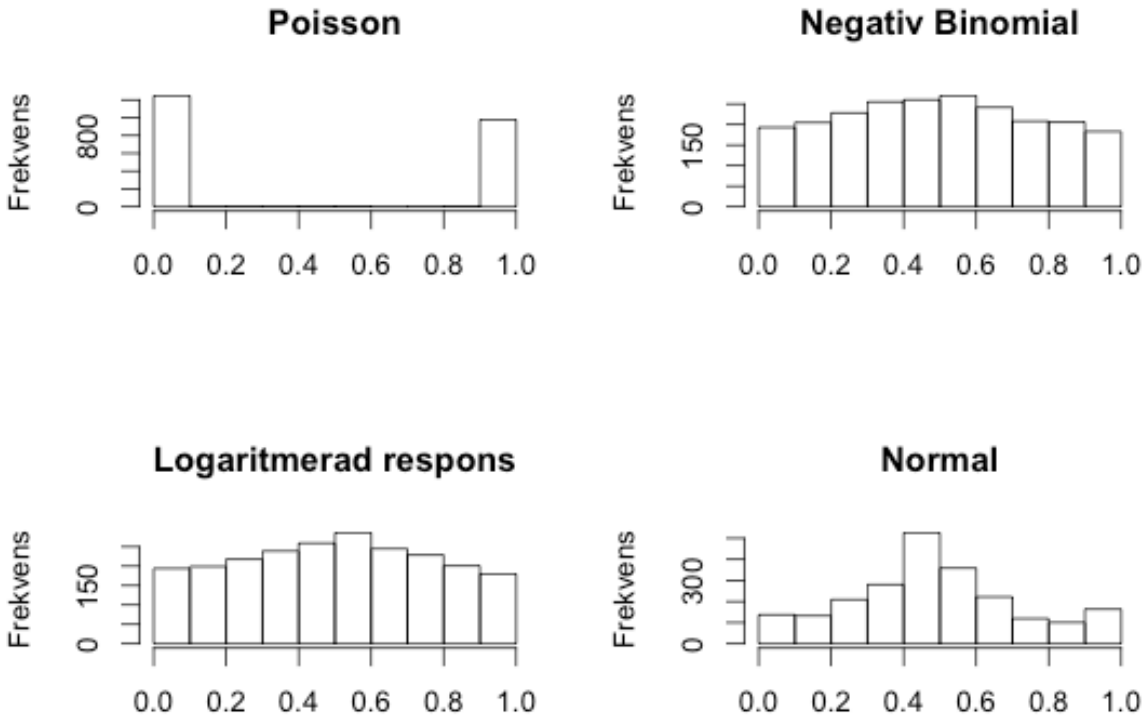
Tabell 5.2. *AIC för olika modeller efter stepAIC*

Modell	**AIC efter stepAIC**
Logaritmerad respons	333
Poisson	47 564 359
Negativ Binomial	56 491
Normal	58 891

4.2 Residualer och PIT

För att undersöka vilken modell som är bäst anpassad till data så tittar vi även på plottar av residualerna för varje modell, dessa finns i Appendix Figur 1- Figur 4. Vi ser tydligt att modellerna Poisson och Normal är de med sämsta residualer. Båda dessa har stora residualer och ingen av modellernas residualer ser ut att följa att följa en normalfördelning i kvantilploten. Både negativ binomialfördelningen och logaritmerad respons har en kvantilplot som ser normalfördelad ut. Det går dock inte att jämföra residualerna mellan dessa två modeller.

I Figur 5.1 nedan presenteras histogram av PIT värden. Då vi begrundar histogrammet för modellen Poisson ser vi att denna fördelning är underspridd. Modellen normal har ett histogram som är mer bergsformat vilket tyder på att denna fördelning är överspridd. Både modellerna logaritmerad respons och negativ binomial har PIT-värden som ger likformigt fördelade histogram. Från histogrammen är det dock svårt att avgöra vilket av dessa som är mest fördelaktigt.



Figur 5.1. *PIT för de modellerna Poisson, Negativ Binomial, Logaritmerad Respons och Normal.*

4.3 Skattningar

I Tabell 5.3 presenteras skattningar för de fyra olika modellerna. Att notera är att två av skattningarna för modellen normalfördelningen inte är signifikanta. Samtliga skattningar för personernas faktorvariabler saknar signifikans i denna modell, dessa presenteras inte i studien. Övrigt att notera är att samtliga skattningar för modellerna Negativ binomialfördelningen och modellen logaritmerad respons ger snarlika värden. Spridningsparametern för negativ binomialförelningen skattas i R till 0,06.

Från Tabell 5.3 ser vi även att parametrarna β_1 och β_2 är de kategoriska variabler som har en positiv skattning, alltså x_1 och x_2 har en positiv inverkan på antalet likes jämfört med basvariabeln x_3. Parametrarna β_4 och β_5 har en negativ skattning, dvs. kategorierna x_4 och x_5 har då en negativ inverkan på antalet likes jämfört med basvariabeln x_3. Det gäller för samtliga modeller att bilder i kategori x_1 ger flest likes och bilder i kategori x_2 näst flest. Bilder från kategori x_5 ger minst likes och bilder i kategori x_4 ger näst minst likes. Parametern β_3 har ingen skattning eftersom att detta är basvariabeln för de kategoriska variablerna. Dess effekt beskrivs enbart av interceptet.

β_6 har ett positiv värde, dvs. motsvarande variabel har en positiv inverkan på antalet likes i samtliga modeller. Vi ser även att parametrarna β_7, β_8 och β_9 samtliga har ett negativt värde, alltså dess variabler inverkar negativt på antalet likes för samtliga modeller.

Tabell 5.3. *Skattningar för slutgiltiga modeller. Värden markerade med * innebär att variabeln inte är signifikant skild från noll på en 10 % signifikansnivå.*

Variabel	**Poisson**	**NegBin**	**Normal**	**Log(Respons)**
Intercept	-5.4730	-5.5040	-5.4428	-5.5359
β_1 Bild på endast personen	0.1498	0.2339	0.0966	0.2343
β_2 Bild på personen och en till	0.0456	0.0832	0.0228*	0.0812
β_3 Bild på personen och flera	Basvariabel	Basvariabel	Basvariabel	Basvariabel
β_4 Bild på andra personer	-0.3169	-0.3275	-0.3144	-0.3306
β_5 Bild utan ansikten	-0.3226	-0.3717	-0.3178	-0.3779
β_6 Hur länge bild har legat ute	0.0014	0.0022	0.0008	0.0023
β_7 Antal hashtags	-0.0079	-0.0134	-	-0.0155
β_8Antal texttaggar	-0.0298	-0.0303	-0.0338	-0.0312
β_9 Antal bildtaggar	-0.0072	-	-0.0135*	-

4.4 Sammanfattning av resultat och slutgiltig modell

Från undersökningen av AIC ser vi att Poissonfördelningen är den modell med sämst anpassning och att negativ binomialfördelningen är den modellen av de med GLM som ger bäst anpassning. Vi ser även att stepAIC inte ger någon reducering av variablerna för Poisson. Från skattningarna ser vi att normalfördelningen ger sämst signifikanser samt att dess skattningar skiljer sig mest från de övriga modellerna. När vi studerar residualerna så noterar vi att både Poisson och normal inte har önskat resultat för residualerna. Negativ binomial och logaritmerad respons däremot har desto bättre.

Skattningar för samtliga personers faktorvariabler presenteras i Tabell 1 i Appendix för modellerna negativ binomialfördelningen och logaritmerad respons. De flesta av dessa skattningar likt parameterskattningarna i Tabell 5.3 är snarlika för de två modellerna. Detta tillsammans med histogrammen i Figur 5.1 indikerar att modellerna är ungefär lika bra anpassade till data.

Eftersom modellen med logaritmerad respons inte kan ta hänsyn till bilder med noll likes samt att denna hanterar kontinuerlig data kommer vi dock välja negativ binomial som den bästa modellen för data.

I Tabell 5.4 presenteras de faktiska värdena för de 20 första observationerna i datamaterialet tillsammans med de predikterade värdena samt dess differenser för modellen negativ binomialfördelningen. I tabellen finns även den procentuella andelen likes för de predikterade värdena av de observerade värdena, där 100 % är önskvärt och innebär att vår modell har predikterat det faktiska värdet. Dessa observationer kommer samtliga från Justin Bieber som varken har minst eller mest följare. Observationerna har valts ut ometodiskt och är avsedda att illustrera variationen inom en person.

Den största negativa skillnaden mellan de faktiska och observerade värdet är observation 9 med en skillnad på -507719, modellen har alltså endast predikterat 62 % av det verkliga antalet likes. Den största positiva skillnaden är från observation 19 på 261022 vilken har predikterat 121 % av det faktiska antalet likes. Observation 3 är dock den observation predikterat procentuellt flest likes på 133 %. Den bästa prediktionen får observation 11 som har en differens på 5533, denna observation har predikterat 100 % av antalet likes. Vi noterar även att observationerna 14 och 15 eller observation 16 och 17 ger exakt samma predikterade värde.

Tabell 5.4. *20 faktiska värden och predikterade värden för Justin Bieber med modellen negativ binomial samt differensen och den procentuella andelen likes som modellen predikterar.*

Observation	Faktiskt värde	Predikterat värde	Differens	Andel
1	1 405 333	1 139 166	-266 167	81%
2	1 277 555	1 243 516	-34 039	97%
3	595 026	792 406	197 380	133%
4	1 019 080	1 211 735	192 655	119%
5	1 383 248	1 106 469	-276 779	80%
6	1 172 926	1 251 730	78 804	107%
7	1 026 021	794 147	-231 874	77%
8	1 197 003	854 077	-342 926	71%
9	1 324 848	817 129	-507 719	62%
10	1 012 272	854 077	-158 195	84%
11	1 491 866	1 497 398	5 532	100%
12	1 058 353	873 027	-185 326	82%
13	1 224 898	1 281 806	56 908	104%
14	1 359 717	1 530 622	170 905	112%
15	1 323 776	1 530 622	206 846	115%
16	1 240 926	1 316 532	75 606	106%
17	1 327 664	1 316 532	-11 132	99%
18	1 061 799	1 175 231	113 432	110%
19	1 229 228	1 490 250	261 022	121%
20	1 282 625	1 426 608	143 983	111%

Då vi inte vet om Justin Bieber av någon anledning skulle ha bättre eller sämre skattningar än någon annan presenteras även den första observationen för varje person i Tabell 2 i appendix. I denna tabell ser vi att den observation som gett överlägset flest likes procentuellt i vår modell jämfört med det faktiska värdet är observationen från Nicole Polizzi som har predikterat 254 % av antalet likes. Den observation som predikterat minst antal likes procentuellt är observationen från Tyga som endast predikterat 51 % av det faktiska antalet likes. I övrigt finns det flera observationer som ligger nära 100 %.

5 Diskussion

5.1 Resultat och skattningar

Från kapitlet resultat kommer vi fram till att negativ binomialfördelningen är den modell som passar bäst för att beskriva data. Detta är alltså svaret på frågan vilken modell som bäst förklarar antalet likes. Då vi studerar Tabell 5.4 med de förväntade värdena vi får från den slutgiltiga modellen mot de faktiska värdena för Justin Bieber ser vi att vissa observationer ligger relativt nära de faktiska värdena och att andra observationer skiljer relativt mycket mot de riktiga värdena. Det beror förmodligen på att bilderna har fler attribut än de som behandlas i denna studie. Vi ser att observationerna 14 och 15 samt observationerna 16 och 17 får exakt samma predikterade värde. Det bör dock inte gå läsaren förbi att de verkliga värdena inte är allt för olika varandra heller. Detta kan beskrivas av den naturliga variation som uppstår,

alternativt att det finns flera variabler som påverkar antalet likes som inte behandlas i denna analys.

I Tabell 2 i appendix ser vi den första observationen för varje person samt dess predikterade värde. Det är dock svårt att dra några slutsatser från denna tabell då vi inte vet hur mycket variation det finns inom varje person. Från Figur 2.1 ser vi även att vissa personer tycks ha större variation än andra. Även då vissa predikterade värden är relativt långt ifrån de verkliga värdena i Tabell 2 i Appendix, lyckas modellen i många fall ge en god indikation på hur många likes en bild kommer generera.

För att besvara frågan hur antalet likes påverkas av vilken typ av bild som visas ser vi i Tabell 5.3 vilka variabler som bidrar till fler eller färre likes för en bild. I modellen är β_5 den parametrar som har minst värde av de kategoriska variablerna på -0.3717. Det betyder att en bild i kategori $x_5 = Inga\ ansikten$ ger minst antal likes av de fem olika kategorierna. Vilket även stöds av slutsatserna från artikeln av Bakhshi, Shamma & Gilbert (2014). Näst minst likes får en bild i kategori $x_4 = Andra\ personer\ utan\ personen\ som$ äger $kontot$ som har en negativ parameterskattning på -0.3275. Efter denna basvariabeln $x_3 = Personen\ som$ äger $kontot\ och\ flera\ personer\ till$ som enbart beskrivs av interceptet. Näst mest likes ger variabeln $x_2 = Personen\ som$ äger $kontot\ och\ en\ person\ till$ som har en parameterskattning på 0.0832. Variabeln $x_1 = Endast\ personen\ som$ äger $kontot$ är den variabel som ger flest antal likes med en parameterskattning på 0.2339.

Av övriga variabler ser vi att β_6 är den enda parameter som har positivt värde, alltså att $x_6 = Hur\ länge\ en\ bild\ har\ legat\ ute$ ger en positiv effekt på antalet likes. Dvs. ju längre en bild ligger ute, desto fler likes genererar den. Från tabellen ser vi även att parametrarna β_7 och β_8 får negativa skattningar. Det innebär att taggar inverkar negativt på antalet likes.

Från Tabell 5.3 och Tabell 7 i Appendix ser vi att den bild som skulle generera flest likes enligt vår modell är en bild utlagd av Harry Styles från One Direction på endast honom, utan någon form av taggar som har legat ut väldigt länge.
Den bild som skulle generera minst likes skulle vara en bild utan ansikten av Akon som är vår bas variabel. Bilden skulle ha legat ute kortast möjliga tid och skulle innehålla maximalt antal texttaggar och hashtags.

5.2 Framtida forskning och brister

De modeller som tas fram i denna studie använder faktorvariabler för de personer som bilderna har samlats ihop ifrån. Det innebär alltså att vi inte kan estimera antalet likes för någon annan person än de som inkluderas i studien. För att kunna göra detta måste vi först hitta skattning för personens faktorvariabel.

Det kan tänkas att de individuella effekter som beskrivs av personernas faktorvariabler istället kan beskrivas att andra faktorer som tillexempel kön, bransch, ålder etc. Det skulle därför vara intressant att undersöka om det går att göra en modell med mer generella variabler än faktorvariablerna för personerna som skulle kunna tillämpas på andra användare.

Det är tänkbart att personernas följare är olika mycket aktiva samt att variationerna är olika beroende på vilken målgrupp kändisen har. Det är inte orimligt att personer med yngre målgrupper har följare som är mer aktiva på sociala medier. Det skulle därför även vara intressant att göra analyser på hur internet beteendet är hos olika individer och vilka faktorer som kan beskriva detta.

Om det finns intresse för någon speciell individ skulle det även kunna vara av intresse att analysera flera av dennes bilder. Vi vet inte om vissa personer får fler likes på någon specifik bild än andra, oberoende av personens faktorvariabel. Studien skulle då kunna utökas till att jämföra färre personer i varsin modell.

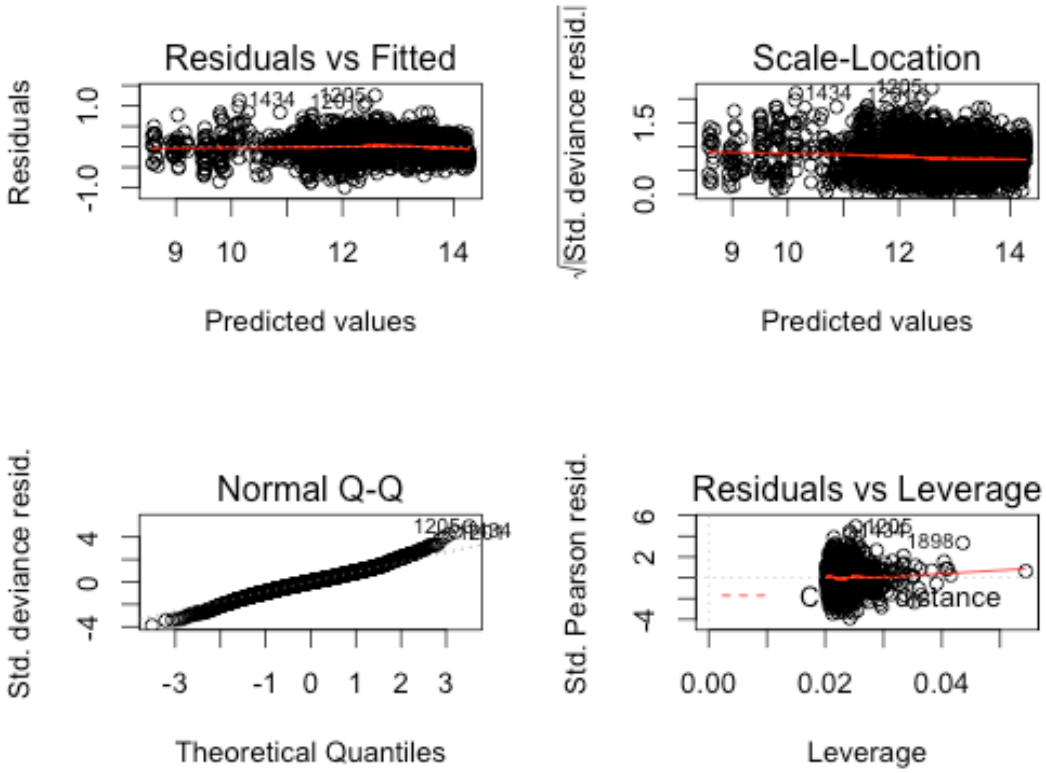
Instagram är ett media som visar nyaste bilder först. Det kan alltså verka tveksamt att variabeln som beskriver hur länge en bild har legat ute skulle ha en effekt överhuvudtaget på antalet likes. Därför skulle det vara intressant att analysera bilder som har legat ute längre än bilderna som finns i vårt datamaterial.

Eftersom att data i studien är tagen endast från kända personer kan det vara intressant att undersöka om dessa skattningar skulle fungera för även okända personer, eller om skattningarna skulle bli annorlunda. Alltså att göra en liknande analys på endast okända personer.

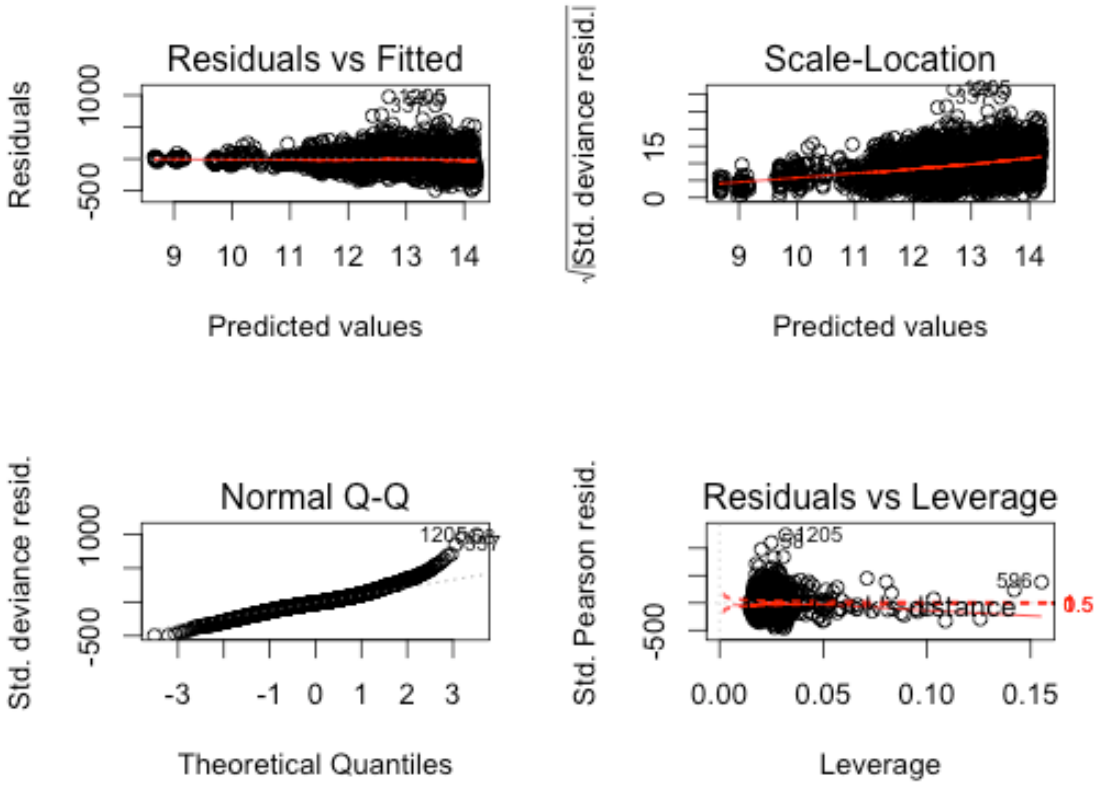
Det går även att göra fler liknande studier där man kollar på fler kategorier för att se exempelvis hur djur, mat etc. påverkar antalet likes. Det skulle även gå att göra analyser på bildtexten och vilken effekt till exempel nyckelord eller användandet av emojis har för antalet likes.

Liknande studier skulle kunna tillämpas på andra sociala medier. Det skulle vara intressant att se om andra medier skulle ge liknande skattningar för de olika kategorierna.

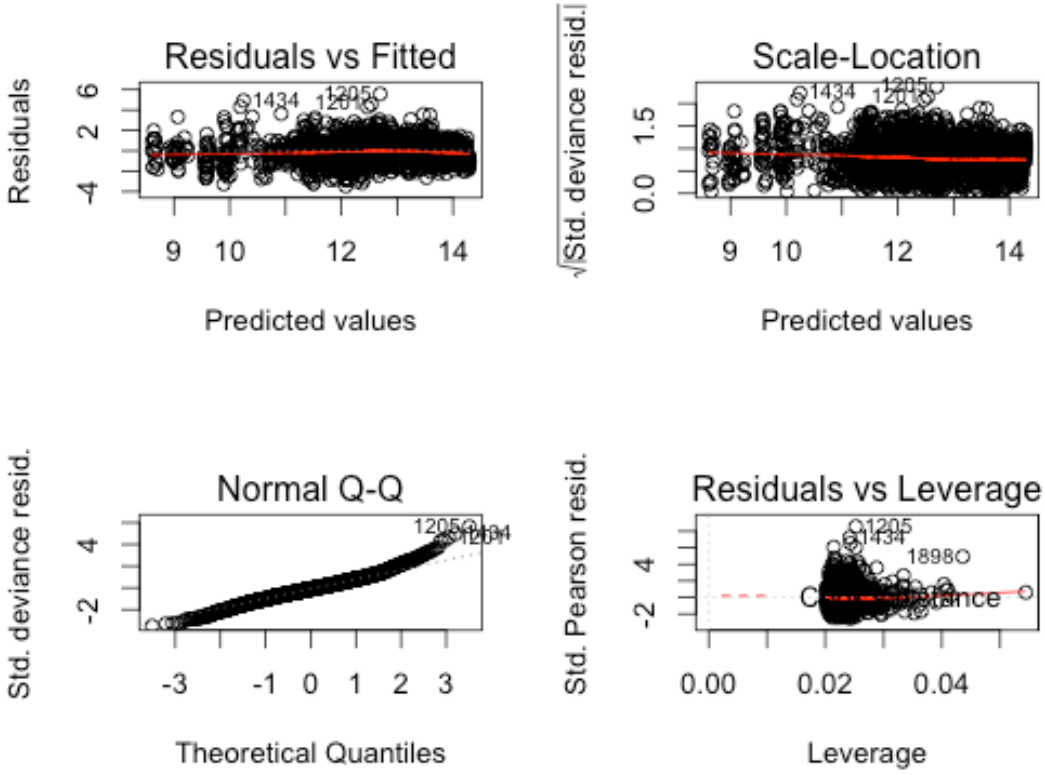
Appendix



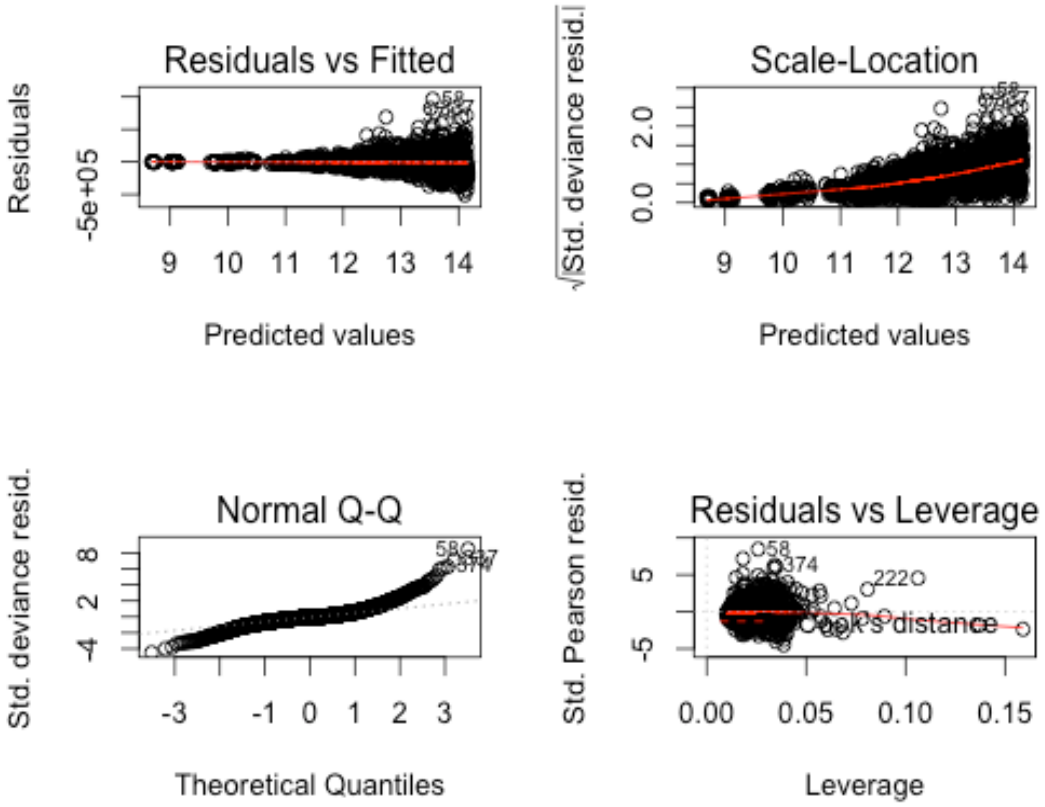
Figur 1. *Diagnostik plot för modellen logaritmerad respons*



Figur 2. *Diagnostik plot för modellen Poisson.*



Figur 3 *Diagnostik plot för modellen Negativ Binomial.*



Figur 4 *Diagnostik plot för modellen Normal.*

Tabell 1. *Skattningar för faktorvariablerna för varje person i modellerna Negativ Binomial och logaritmerad respons.*

Person	**Nr**	**Antal Följare**	**β för NegBin**	**β för Log(Respons)**
Justin Bieber	1	22 000 000	2.539057	2.5495392
Kim Kardashian	2	25 000 000	1.971053	1.9539373
Rihanna	3	14 000 000	2.031400	2.0160420
Beyonce	4	25 000 000	2.019768	2.0169426
Ariana Grande	5	24 000 000	2.418571	2.4304168
Miley Cyrus	6	15 000 000	1.635725	1.6259036
Kendall Jenner	7	18 000 000	2.544243	2.5391059
Taylor Swift	8	21 000 000	2.569174	2.5704001
Khloé Kardashian	9	17 000 000	2.109988	2.1079226
Kylie Jenner	10	17 000 000	2.744015	2.7584879
Kourtney Kardashian	11	14 000 000	2.044729	2.0561358
Selena Gomez	12	23 000 000	2.483099	2.4999426
Harry Styles	13	10 000 000	3.396635	3.4198375
Kevin Hart	14	11 900 000	1.660752	1.6607982
Niall Horan	15	8 700 000	3.349169	3.3652171
Jenna Marbles	16	3 100 000	2.354008	2.3696385
Lebron James	17	8 800 000	1.859435	1.8158143
Nicole Polizzi	18	5 500 000	1.810609	1.7729243
Akon	19	1 900 000	Basvariabel	Basvariabel
Tyga	20	4 700 000	1.792303	1.7846055
Nicki Minaj	21	16 300 000	1.678512	1.6924643
Cara Delevingne	22	9 700 000	2.340405	2.3557361
Demi Lovato	23	11 600 000	2.543038	2.5592430
Neymar Júnior	24	15 700 000	2.258300	2.2803024
Justin Timerlake	25	7 900 000	2.072258	2.0007276
Katy Perry	26	15 100 000	2.041409	2.0528383
Cristiano Ronaldo	27	12 700 000	2.627724	2.6459911
Ashley Benson	28	5 700 000	2.388443	2.3987401
Sean Diddy Combs	29	4 000 000	0.622251	0.5587636
Miranda Kerr	30	5 500 000	2.004585	2.0041501
Austin Mahone	31	5 700 000	2.421987	2.4446276
Schott Disick	32	7 100 000	2.157235	2.1681453
Lucy Hale	33	6 200 000	2.367613	2.3715193
Channing Tatum	34	4 800 000	2.354078	2.3407672
Lady Gaga	35	5 900 000	2.123501	2.1394621
Lauren Conrad	36	3 900 000	2.121367	2.1184548
Shay Mitchell	37	5 300 000	2.591191	2.6078753
Jessica Alba	38	4 300 000	1.300501	1.3019899
Zendaya Maree	39	5 300 000	2.450962	2.4662460
Wiz Khalifa	40	5 300 000	2.064242	2.0772162
Snoop Dogg	41	4 300 000	0.185844	0.1554105
The Rock	42	7 600 000	2.045104	2.0715383
Dwyane Wade	43	3 700 000	1.607261	1.6238625
Macklemore	44	3 000 000	2.137492	2.1496920
Zooey Deschanel	45	3 300 000	2.466102	2.4671443

Tabell 2. *Den första observationen, predikterade värdet, differensen och procentuella andelen predikterade likes mot observerade likes för varje person med modellen negativ binomial.*

Person	Antal följare	Obs värde	Predvärde	Differens	Andel
Justin Bieber	22 000 000	1 405 333	1 139 166	-266 167	81%
Kim Kardashian	25 000 000	573 955	497 983	-75 961	86%
Rihanna	14 000 000	383 923	467 922	83 999	121%
Beyonce	25 000 000	1 164 321	846 260	-318 061	73%
Ariana Grande	24 000 000	622 345	725 589	103 244	116%
Miley Cyrus	15 000 000	275 473	210 544	-64 929	76%
Kendall Jenner	18 000 000	834 602	924 454	89 852	110%
Taylor Swift	21 000 000	1 182 196	1 415 985	233 789	119%
Khloé Kardashian	17 000 000	652 174	413 072	-239 102	63%
Kylie Jenner	17 000 000	564 170	745 030	180 860	132%
Kourtney Kardashian	14 000 000	230 301	288 004	57 703	125%
Selena Gomez	23 000 000	1 271 934	1 223 907	-48 027	96%
Harry Styles	10 000 000	778 172	818 385	40 213	105%
Kevin Hart	11 900 000	293 025	266 723	-26 302	91%
Niall Horan	8 700 000	657 518	676 016	18 498	102%
Jenna Marbles	3 100 000	101 603	114 409	12 806	112%
Lebron James	8 800 000	220 129	237 078	16 949	107%
Nicole Polizzi	5 500 000	38 886	99 065	60 179	254%
Akon	1 900 000	10 597	10 209	-388	96%
Tyga	4 700 000	163 167	83 120	-80 047	51%
Nicki Minaj	16 300 000	589 082	450 054	-139 028	76%
Cara Delevingne	9 700 000	422 133	434 182	12 049	102%
Demi Lovato	11 600 000	860 418	734 262	-126 156	85%
Neymar Júnior	15 700 000	678 995	775 761	96 766	114%
Justin Timerlake	7 900 000	476 908	323 310	-153 598	68%
Katy Perry	15 100 000	789 338	604 602	-184 736	77%
Cristiano Ronaldo	12 700 000	1 038 103	909 964	-128 139	87%
Ashley Benson	5 700 000	185 396	275 922	90 526	148%
Sean Diddy Combs	4 000 000	11 107	20 064	8 957	180%
Miranda Kerr	5 500 000	65 326	110 155	44 829	168%
Austin Mahone	5 700 000	241 330	286 590	45 260	118%
Schott Disick	7 100 000	169 838	171 118	1 280	101%
Lucy Hale	6 200 000	402 441	269 865	-132 576	67%
Channing Tatum	4 800 000	92 284	141 784	49 500	153%
Lady Gaga	5 900 000	253 718	254 206	488	101%
Lauren Conrad	3 900 000	90 935	89 358	-1 577	98%
Shay Mitchell	5 300 000	311 294	198 919	-112 375	64%
Jessica Alba	4 300 000	32 905	44 573	11 668	135%
Zendaya Maree	5 300 000	175 869	181 107	5 238	103%
Wiz Khalifa	5 300 000	116 719	170 325	53 606	145%
Snoop Dogg	4 300 000	30 219	26 686	-3 533	88%
The Rock	7 600 000	228 188	287 630	59 442	126%
Dwyane Wade	3 700 000	66 578	52 704	-13 874	79%
Macklemore	3 000 000	88 417	72 477	-15 940	82%
Zooey Deschanel	3 300 000	140 402	177 216	36 814	126%

Referenser

Agresti A. (2002). *Categorical Data Analysis*. John Wiley & Sons Inc. Hoboken, New Jersey.

Andersson P. och Tyrcha J. (2014). *Notes in Econometrics*. Matematiska Institutionen, Stockholms universitet.

Aragan, J., D. Eberly and S. Eberly. (1991) *Existence and Uniqueness of the Maximum Likelihood Estimator for the Two-parameter Negative Binomial Distribution*. Department of statistics, Stanford univeristy. Stanford. No 282.

Bakhshi S, Shamma D. A. och Gilbert E. (2014) *Faces engage us: Photos with faces attract more likes and comments on instagram*. URL: http://www.cc.gatech.edu/~sbakhshi/faces2014-bakhshi.pdf

Bakhshi S., Gilbert E. , Chang S. och Terveen L. (2013) *"I need to try this!": A statistical overview of Pinterest*. URL: http://www.cc.gatech.edu/~sbakhshi/pinterest2013-gilbert.pdf

Bondarouk T och Olivas-Luján R. M. (2013). *Social Media in Strategic Management*. Emerald Group Publishing Limited. Bingley.

Czado C. Gneiting T. och Held L. (2009). *Predictive model assessment for count data*. Biometrics no. 65, s 1254 - 1261.

Fox J. (2015) *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, Inc. United States.

Liero H. Och Zwanzig S. (2012). *Introduction to the Theory of Statistical Inference*. CRC Press, Taylor & Francis Group. Boca Raton.

Roosevelt H.Q. (2014). *Hämtat februari 2015. The 50 most popular people on Instagram in 2014*. The Roosevelts. URL: http://www.rsvlts.com/2014/01/07/popular-people-instagram-2014/

Greener S. och Rospigliosi A. (2014). *Proceedings of the European Conference on Social Media*. Academic Conferences and Publishing International Limited. UK.

R Core Team (2015). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wolf M. (2013). Hämtat februari 2015. *Predicting the Number of Likes on a facebook status With statistical Keyword Analysis*. Minimaxir.com. URL: http://minimaxir.com/2013/06/big-social-data/

Wolf M. (2015). Hämtat februari 2015. *Locating All the Christmas Trees on Instagram*. Minimaxir.com. URL: http://minimaxir.com/2015/01/tree-time/.