Kunskapskontroll R

Daniel Borgenstedt

EC Utbildning

202404

Innehåll

1 Inledning

I denna rapport är tanken att prediktera statistik, med programmeringsspråket R, från SCB som täcker in olika YH-utbildningar och att försöka förutspå antal elever som tar examen utifrån resultatet av tidigare elever som påbörjat en YH-utbildning. För att uppfylla syftet så kommer följande frågeställning att besvaras:

Går det att skapa en linjär modell som kan prediktera antal elever som tar examen från en YH-utbildning?

I en del av kunskapskontrollen ingick datainsamling i grupp som kommer behandlas under separat stycke. Därefter kommer sju utvalda frågor att besvaras längst bak. Rapporten ämnar täcka in betyget Godkänt.

2 Teori

2.1 Modeller och dataset

Linear Model

Linjära modeller används när variabeln är kontinuerlig och förväntas ha ett linjärt samband med ”förklarings”-variablerna som kan vara en eller flera stycken (geeksforgeeks.org, 2023).

Generalized Linear Model

GLM består utav en del regressionsmodeller. De används till att upptäcka relationer mellan respons-variabler och en eller flera prediktorer. Till skillnad från vanliga linjära regressions-modeller som antar ett linjärt förhållande till respons och prediktor-variabler, så kan GLM tillåta icke-linjära förhållanden (geeksforgeeks.org, 2023).

Dataset

Tabellen med statistiken är hämtad från SCB och innefattar yrkeshögskoleprogram från åren 2012–2023. Endast antagna elever finns med under 2023, då läsåret är pågående. Datasetet innehåller 300 celler uppdelat på 12 rader och 25 kolumner. Programmen som tabellen innefattar är:

Företagsförsäljning
Redovisningsekonom
Systemhantering och programmering
Webbutvecklare
Hotell Management
Yrkessvetsare
Fordonstekniker
Bagare/konditor
Kart- och mättekniker
Byggledare
Fastighetsförvaltare
Djurvårdare

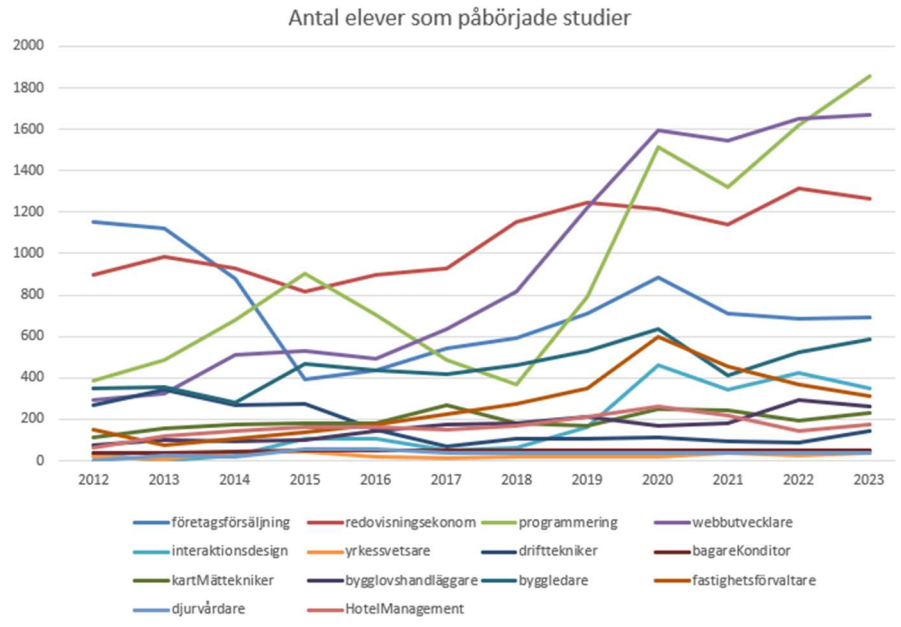
3 Metod

3.1 Databas

Statistiska Centralbyrån – SCB, tillhandahåller statistik inom olika ämnen och årtal som är öppen för allmänheten att ta del av. Data för diverse Yrkeshögskole-utbildningar från åren 2012 - 2023 togs ner i tabellformat. Antal antagna som påbörjat studier avser personer som varit studerande tre veckor efter utbildningens start. Antalet studerande avser aktiva studerande på utbildningsomgångar som pågått minst en dag under en avsedd period. Jag har valt att fokusera på antal antagna då det sträcker sig över åtminstone tre veckor.

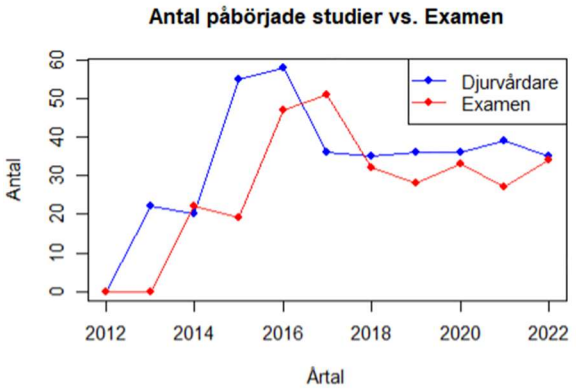
3.2 Upplägg

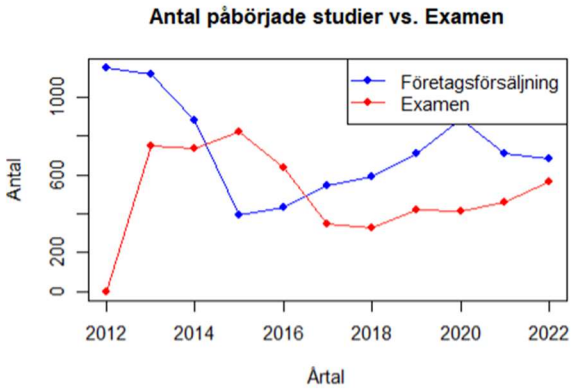
Urvalet gjordes på samtliga rader av tabellen utom den sista då den raden var tänkt att predikteras utifrån de senaste årens siffror gällande antal studerande som påbörjat ett YH-program. För att begränsa urvalet valdes 12 st. YH-program ut från SCB.



3.4 Vald modell

Jag undersökte främst Linear Model och Generalized Linear Model eftersom sambandet vid första översikt antogs vara linjärt. I slutändan blev det GLM. Som parametrar fokuserades det främst på Poisson.





Tyvärr visade det sig vara mer invecklat än att man kunde göra en slutsats via linjärt samband. Detta då vissa år hade flera examinerade än vad som var elever som påbörjat studierna tidigare år.

3.5 Outliers och avvikelser

Vid närmare granskning upptäcktes flera celler med väldigt hög siffra för examinerade jämfört med det antal som hade påbörjat studierna samma år. Det förklarades av att datan för examinerade inkluderade tidigare årgångars studenter som gjort kompletteringar för att kunna få ut sin examen.

Då dessa gjordes samma år som efterkommande elever läggs dessa till i statistiken för samma år. På grund av detta blev prediktionerna väldigt orealistiska. Vissa program med ett fåtal elever fick prediktion på en avsevärd högre siffra för elever som skulle ta examen.

Examinerade avser antagna som har uppfyllt alla villkor för examen. Examinerade läggs därefter till det slutår som en utbildningsomgång har. För examinerade finns därför en eftersläpning i statistiken på grund av sena kompletteringar. Uppgifter för det senaste referensåret redovisas därför i november.

Det står inte heller hur lång YH-utbildningen är. Då de flesta dock är två år, så finns det många som endast är ett år och även någon enstaka som är 2,5 år. En utbildning kan även ändras och förlängas från ett år till ett annat vilket kan försvåra en analys avsevärt.

Om man förutsätter att föregående år tar examen efterkommande år som minsta möjliga längd, så blir prediktion gjord utifrån föregående rad i tabellen, vilket borde göra att sista raden teoretiskt sett kan predikteras.

Då en generell modell för alla inte fungerade bra, så kom jag till slutsatsen att jag borde försöka isolera dessa problem. För att göra det, så då gjordes separata modeller på respektive program. Detta gjordes med en loop som gick igenom samtliga kolumnpar som bestod av antal elever som påbörjat en utbildning med antal som tagit ut examen för samma. Trots isolering per utbildningskolumn, så visade det sig att resultatet inte blev förbättrat i slutändan. Jag prövade både LM och GLM men det var inte någon märkbar skillnad.

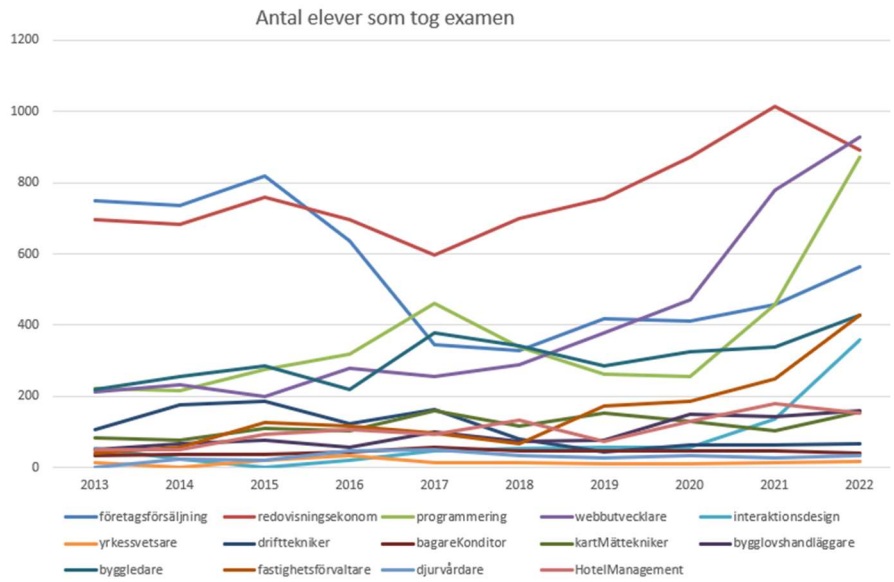
Resultat och Diskussion

4.1 Sammanfattning

Enligt statistik från SCB kring ” Studerande och examinerade i yrkeshögskolan efter utbildningens inriktning, tabellinnehåll och år”, så sticker vissa ämneskategorier ut i att de har högre procent som tar examen.

Till dessa tillhör ”Kultur, media och design” samt ”Lantbruk, djurvård, trädgård, skog och fiske” där andel examinerade uppgick till ~82-83%. Andra kategorier som ”Data/IT”, ”Ekonomi, administration och försäljning”, ”Hotell, restaurang och turism”, ”Samhällsbyggnad och byggteknik” samt ”Teknik och tillverkning”, har en examensgrad på ca ~58-61%. (SCB, 2024).

Ekonomi och IT-utbildningar är utbildningar som ökar stadigt i popularitet för varje år, dock ser man ingen större skillnad i hur många som examinerar sig. Det verkar vara liknande resultat år efter år. Om detta beror på svårighetsgraden på utbildningen eller andra faktorer är en fråga för sig.

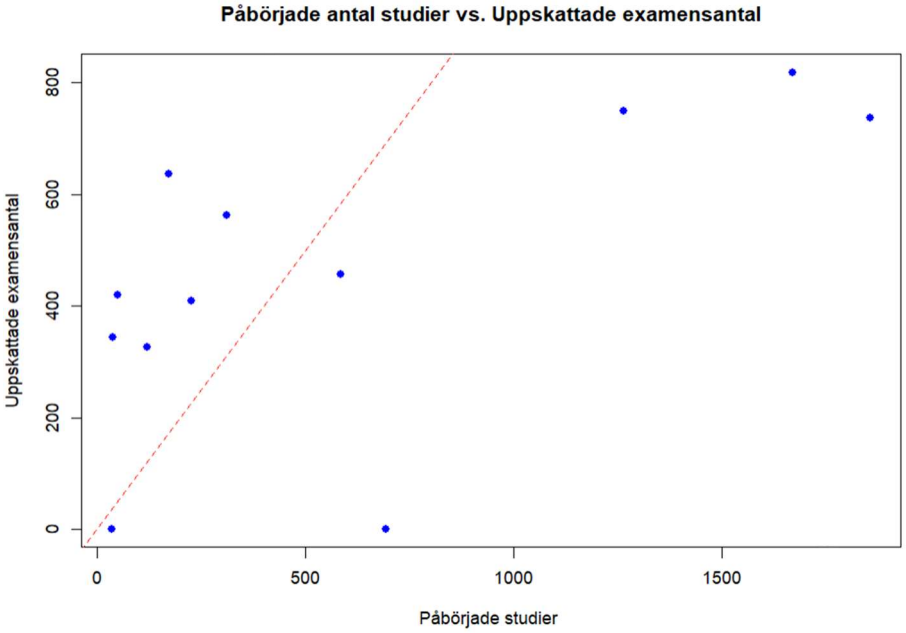


5 Slutsatser

Går det att skapa en linjär modell som kan prediktera antal elever som tar examen från en YH-utbildning?

I detta fallet verkar det inte gå att göra det tillförlitligt. Anledningarna till det är flera. Datasetet kan mycket väl ha varit för litet för att det ska ha gått att göra en bra modell. Men även det faktum att man inte benämner längden på en utbildning spelar in. Utbildningen kan vara allt från 1 år till 2,5 år. Lägg därtill de elever som tar en kompletterande examen efterkommande år, gör att statistiken kan skilja sig avsevärt med över hundra från det antal som har påbörjat en utbildning. Jag trodde att isoleringen av respektive program med en separat modell skulle ge någon form av förbättring men

det verkar inte ge någon större skillnad. Hade det funnits mer detaljerad information kring varje år så tror jag att det funnits ett tydligare linjärt samband, men i denna tabell, så blir statistiken omblandad pga. hur man räknar elever som tagit examen och det gör att det kanske inte alltid stämmer med egentliga siffror. Följande illustrerar den sista raden som predikteras utifrån föregående års celler.



6 Datainsamling

Under kunskapskontrollen utfördes en datainsamling i grupp kring bilförsäljning på Blocket. I min grupp ingick, Abdulrahman, Alia, Anton, George, Goran, Jesper, John och Kawser. Till en början var det lite rörigt, men jag tog på mig en guidande roll och ställde frågor och gav förslag som gjorde att vi tog oss framåt. Som i alla grupper är det vissa som pratar mer än andra och man får försöka dra in alla i någon form för att det ska vara mer som en gruppuppgift. Vi la därför in att alla skulle hämta in ett mindre antal bilar av ett märke för att vi skulle kunna göra en minimodell på någon statistisk teori och jämföra.

Jag gjorde en s.k. ”webscraping” från Blocket och tog ner ca 7000 bilannonser som vi i gruppen kunde sammanställa på eget håll och jämföra med minimodellen. En intressant aspekt var att man kunde se en viss prispåverkan kring vilken typ av bränsle det var i en bil, t.ex. en Volvo V40, och hur det påverkade prisbilden.

	Brand	Model	ModelYear	Engine	Miles	gears	Price	Region	Dealer
	<chr>	<chr>	<dbl>	<chr>	<dbl>	<chr>	<dbl>	<chr>	<chr>
1	Volvo	V40	2015	Bensin	8447	Manuell	174900	Stockholm	Bilia Outlet Kista
2	Volvo	V40	2015	Bensin	8270	Automat	199900	Stockholm	Volvo Car Kungsängen
3	Volvo	V40	2015	Bensin	6845	Automat	219800	Stockholm	Aftén Bil outlet Vallentuna
4	Volvo	V40	2015	Bensin	10555	Manuell	144900	Göteborg	KGJ Bil AB
5	Volvo	V40	2015	Diesel	21063	Automat	164800	Stockholm	Riddermark bil Nacka
6	Volvo	V40	2015	Diesel	14400	Automat	164900	Stockholm	Sigtuna bil
7	Volvo	V40	2015	Diesel	7400	Automat	159900	Stockholm	Louise Bilsalong
8	Volvo	V40	2015	Diesel	14840	Manuell	120000	Stockholm	Autohero
9	Volvo	V40	2015	Diesel	12698	Manuell	139900	Stockholm	NA
0	Volvo	V40	2015	Bensin	16800	Manuell	110000	Stockholm	NA

7 Teoretiska frågor

1.Beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

SVAR: Som jag tolkar det, utifrån en normalfördelning, så kommer de olika punkterna att ligga utefter en rät linje när man använder QQ-plot. Om du vill se om en fördelning verkligen följer en normalfördelning så kan man använda det för att visualisera det lättare. Om punkterna inte följer linjen så antyder det att datan inte riktigt efterföljer normalfördelningen.

2. Din kollega Karin frågar dig följande: *”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?”* Vad svarar du Karin?

SVAR: Med prediktion i maskininlärning kan du förutsäga t.ex. en prisutveckling eller om någon tar examen utifrån betyg, med hjälp av en modell (matematiskt), men med statistisk inferens tittar man kanske närmare på orsaker och varför. I ML så är det mer inriktat på att hitta mönster snarare än varför.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

SVAR: Konfidensintervall används väl mer utifrån att man med hög säkerhet vill förutspå något, t.ex med 95% säkerhet. Till exempel snittlön utifrån ålder. Prediktion är mer för att man tror att det kommer hamna mellan A och B.

4. Den multipla linjära regressionsmodellen kan skrivas som:
$Y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_p x_p + \varepsilon$. Hur tolkas beta parametrarna?

SVAR: De är koefficienter. Till exempel β_0 = Interceptet, β_1= lutningen för rät linje.

5. Din kollega Hassan frågar dig följande: *”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?”* Vad svarar du Hassan?

SVAR: Förr använde man inte datorkraft på samma sätt som man gör idag till uträkningar. En metod som BIC kan användas till att lättare balansera ner en modell så att den inte blir för komplex. Det blir mer utav att en modell presterar bra generellt sett istället för att man kör validering och test på den.

6. Förklara algoritmen nedan för ”Best subset selection”

SVAR: Det är ett sätt för att få fram bästa möjliga modell. Börja från noll och lägg på variabler/prediktorer. Minsta möjliga RSS (Residual sum of squares) indikerar vilken modell som passar bäst.

7. Ett citat från statistikern George Box är: *“All models are wrong, some are useful.”* Förklara vad som menas med det citatet.

SVAR: Antar att han menar att en modell inte har svaren på allt dvs det kan finnas dolda orsaker och faktorer som påverkar, men att en del ändå kan prediktera bra.

Appendix A

Slutgiltig modells prediktioner:

```
Program: foretagsforsaljning:
 Påbörjade studier : 693, Uppskattade examensantal: 0, MAE: 693
Program: redovisningsekonom:
 Påbörjade studier : 1263, Uppskattade examensantal: 750, MAE: 513
Program: programmering:
 Påbörjade studier : 1855, Uppskattade examensantal: 737, MAE: 1118
Program: webbutvecklare:
 Påbörjade studier : 1670, Uppskattade examensantal: 819, MAE: 851
Program: HotelManagement:
 Påbörjade studier : 172, Uppskattade examensantal: 637, MAE: 465
Program: yrkessvetsare:
 Påbörjade studier : 38, Uppskattade examensantal: 344, MAE: 306
Program: fordonstekniker:
 Påbörjade studier : 121, Uppskattade examensantal: 327, MAE: 206
Program: bagarekonditor:
 Påbörjade studier : 50, Uppskattade examensantal: 419, MAE: 369
Program: kartmattekniker:
 Påbörjade studier : 228, Uppskattade examensantal: 410, MAE: 182
Program: byggledare:
 Påbörjade studier : 586, Uppskattade examensantal: 457, MAE: 129
Program: fastighetsforvaltare:
 Påbörjade studier : 311, Uppskattade examensantal: 562, MAE: 251
Program: djurvardare:
 Påbörjade studier : 36, Uppskattade examensantal: 0, MAE: 36
```

Källförteckning

geeksforgeeks.org, 2023 - https://www.geeksforgeeks.org/generalized-linear-models/

geeksforgeeks.org, 2023 - https://www.geeksforgeeks.org/ml-linear-regression/#what-is-linear-regression

SCB, 2024 - https://www.statistikdatabasen.scb.se