Kunskapskontroll - R

Undersökning om korrelationen mellan alkoholvanor och trafikolyckor



Mikaela Bennshagen

EC Utbildning

202404

# Abstract

I chose to investigate the correlation between alcohol consumption and fatal traffic accidents in Sweden in the 21st century. I collected data from Transportstyrelsen and Statistiska Centralbyrån and made a regression analysis in order to see if there was a correlation between these two sets of data. I made the analysis using R in R Studio and formatted the data in Excel.

Innehållsförteckning

[Abstract 2](#_Toc165041940)

[1 Inledning 1](#_Toc165041941)

[2 Teori 2](#_Toc165041942)

[2.1 Regressionsmodeller: Linjär och logistisk regression 2](#_Toc165041943)

[2.1.1 Linjär regression 2](#_Toc165041944)

[2.1.2 Logistisk regression 2](#_Toc165041945)

[3 Metod 3](#_Toc165041946)

[4 Resultat och Diskussion 4](#_Toc165041947)

[5 Slutsatser 5](#_Toc165041948)

[6 Teoretiska frågor 6](#_Toc165041949)

[7 Självutvärdering 8](#_Toc165041950)

[Källförteckning 9](#_Toc165041951)

# Inledning

Jag valde att undersöka om det fanns ett statistiskt samband mellan alkoholkonsumtion och antal omkomna i trafikolyckor. Transportstyrelsen har länge haft en så kallad *nollvision*, som Riksdagen beslutade 1997, som innebär att ingen ska dödas eller skadas allvarligt i trafiken. Då många av de trafikolyckor som sker är på grund av påverkade förare, så ville jag se genom den datan jag samlat, hur de alkoholvanor som finns i Sverige har påverkat denna vision positivt eller negativt.

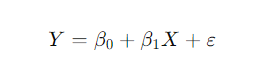
# Teori

## Regressionsmodeller: Linjär och logistisk regression

De två modeller jag valde var linjär regression och logistisk regression. Båda modellerna är vanliga och populära modeller att använda i statistisk analys då de har flera fördelar, inklusive de jag nämner mer utförligt nedan.

### Linjär regression

I linjär regression har man en eller flera oberoende variabler och en kontinuerlig variabel med målet att se om det finns ett linjärt samband mellan variablerna.



Anledningen till att jag valde att ha med denna modell i min undersökning var för att jag tyckte den var lämplig då den har en oberoende variabel (alkoholkonsumtion) och en kontinuerlig variabel (antalet omkomna) vilket passande undersökningens syfte.

### Logistisk regression

Den andra modellen jag valde var logistisk regression. I logistisk regression räknar man ut sannolikheten för en binär (endast två möjliga utfall) beroende variabel som utgår från en eller flera oberoende variabler.



Jag ansåg att logistisk regression var en passande modell i min analys då min undersökning utgår från den binära beroende variabel (omkomna/inte omkomna) baserat på den oberoende variabel (alkoholkonsumtionen). Då antalet omkomna i trafiken är något man vill förhindra så är det lämpligt att använda en modell som gör prediktioner baserat på sambandet mellan omkomna i trafiken och alkoholvanor i Sverige.

# Metod

Jag gjorde min första datainsamling från Transportstyrelsen där jag laddat ner en Excel-tabell som visade antalet omkomna i trafiken sedan år 2003 till år 2023. Min andra insamling av data gjorde jag från Statistiska centralbyrån (SCB) genom att ladda ner från hemsidan till en CSV-fil. Denna CSV-fil innehöll en tabell som visade på en procentandel av riskkonsumenter av alkohol per ålder och kön, som jag sedan formaterade i Excel. Sedan läste jag in båda filerna (och just den fliken i filen från Transportstyrelsen som innehåll antalet omkomna) i R studio där jag sedan utförde mina regressionsanalyser - linjär och logistisk regression - för att få mina resultat.

# Resultat och Diskussion

Resultatet av denna undersökning blev ofullständig då det blev problem med flera element i undersökningen. Jag försökte på olika sätt ladda ner datan till R Studio – genom både webbskrapning, API och andra sätt genom att inspektera hemsidorna. Här stötte jag på många svårigheter oavsett vilken taktik jag körde på bland annat beroende på att vissa hemsidor hade vissa skydd för att försvåra datainsamling. Även om jag till slut lyckades samla in alla data från både Transportstyrelsen och SCB fick jag senare problem med att läsa in datan i R Studio.

Även om jag sedan i R Studio körde korrekta koduppsättningar så uppstod nya problem konstant, oavsett vilken kod jag testade, som gjorde datainläsningen inte blev tillgänglig i R Studio. Jag avslutade med att i vilket fall skriva koden som den potentiellt hade kunnat se ut, utan att köra den.

Förbättringar jag hade kunnat göra, förutom den uppenbara förbättringen av att faktiskt lyckas med koden, är att fördjupa mig ännu mer i skillnader mellan demografiska grupper (kön, ålder etc) vad gäller riskkonsumtion av alkohol. Detta tror jag är möjligt med mer tid, mer arbete och bättre kodstruktur.





# Slutsatser

Om man trots bristerna försöker dra en slutsats av min data och undersökning, så kan vi ändå se att sen början av 2000-talet så har antalet omkomna i trafiken minskat och vi kan samtidigt se en procentuell minskning av andelen riskkonsumenter. Alltså är det rimligt att påstå med viss säkerhet att det finns en korrelation mellan antalet omkomna i trafiken och alkoholkonsumtionen i Sverige och att våra minskade alkoholvanor har lett till en säkrare trafikmiljö.

# Teoretiska frågor

1. Kolla på följande video: https://www.youtube.com/watch?v=X9\_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

**Svar:** Quantile-Quantile metoden är ett sätt att undersöka om ett set med värden är normalfördelade. Metoden går ut på att man ordnar sina värden från lägst till högst och sedan räknar man ut för var och en av de värdena en siffra genom formeln **”i/n+1”** där **”n”** är antalet värden och där **”i”** är **”1”** för det lägsta värdet, **”2”** för det näst lägsta osv. Om värdena är normalfördelade blir QQ en (ungefär) rak linje.

1. 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:** I maskininlärning är fokus på prediktionerna men i statistisk regressionsanalys vill man veta mer i detalj om hur man kom fram till prediktionerna, man vill alltså göra ”the black box” genomskinlig. Om man exempelvis har fem variabler som predikterar ett förväntat huspris är det en enkel prediktion, men om man undersöker hur de fem enskilda variablerna påverkar huspriset i fråga så handlar det om statistisk inferens.

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

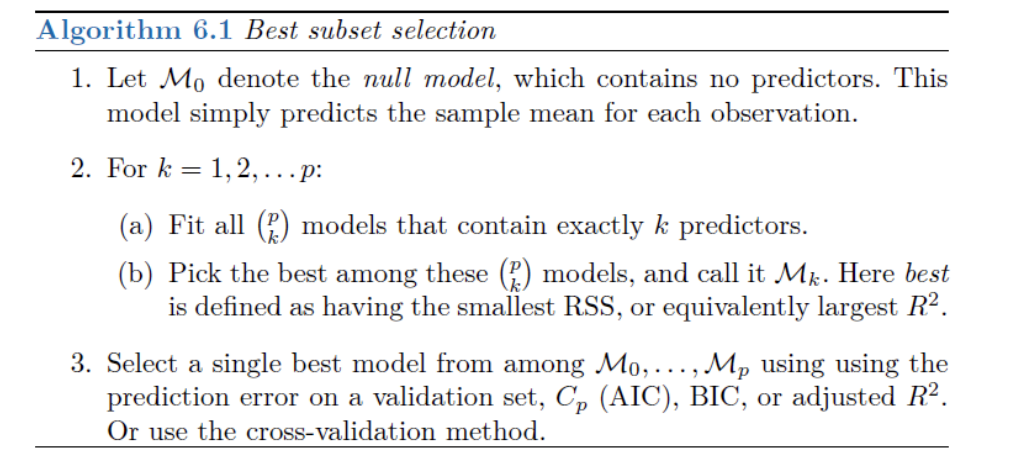
**Svar:** Konfidensintervall handlar om med hur stor säkerhet man kan förutsäga det genomsnittliga värdet på en mängd data (exempelvis på försäljningssiffror). Prediktionsintervall gör samma men för ett specifikt område och därför blir intervallen väsentligt bredare. Exempelvis kan ett konfidensintervall för hela Sverige vara 28-32 men för specifikt Malmö kan prediktionsintervallet vara 24-44.

1. Den multipla linjära regressionsmodellen kan skrivas som: 𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 . Hur tolkas beta parametrarna?

**Svar:** Beta parametrarna är koefficienter som vi tolkar som den genomsnittliga effekten som prediktionerna har på resultatet **(Y)** om vi ökar **x** med en enhet.

1. 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:** Ja det stämmer. Genom att använda mått såsom BIC, validation-set och cross-validation får vi ungefär samma resultat. När då olika modeller ger ungefär lika bra resultat är det logiskt att använda den modell som har lägst antal prediktorer, det vill säga den enklaste modellen.

1. Förklara algoritmen nedan för ”Best subset selection”.

**Svar:** Punkt 1. beskriver att M0 används för ”the null model” när vi inte har några prediktorer utan bara skattar det genomsnittliga värdet för varje observation i urvalet. Punkt 2. Går vi vidare och tittar på alla modeller med 1 prediktion som vi kallar för M1 osv tills vi tittat på alla de bästa (bäst = lägst RSS eller högst R2) prediktionerna för samtliga värden som vi har (M2, M3…Mk). Till sist i punkt 3. väljer vi den bästa enskilda modellen som hittar genom att räkna ut ”the prediction error” på någon av de metoder som nämnts tidigare (BIC, cross- validation etc).

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

**Svar:** Ingen modell är perfekt, det finns alltid brister och osäkerhet. Men en bra modell kan ge en bra skattning. Vissa modeller är bättre än andra och de är användbara när man undersöker något – men man ska bara inte förlita sig för mycket på dem.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

**Svar:** De utmaningar jag haft är de jag redogjort för ovan.

1. Vilket betyg du anser att du skall ha och varför.

**Svar:** Jag hoppas på godkänt.

1. Något du vill lyfta fram till Antonio?

**Svar:** Det är svårt att helt till fullo visa i min rapport hur hårt jag jobbat med denna uppgift och att trots bristande resultat - förutom att ha lärt mig oerhört mycket - också förstått de grundläggande principerna i R programmering.

# Källförteckning

James, G, Witten, D, Hastie, T & Tibshirani, R (2023). *An Introduction to Statistical Learning with Applications in R*. (Second Edition) [PDF]

Prgomet, A (2024) Föreläsningar vid EC Utbildning