Regressionanalys

Teoretiska frågor, grupparbete och självutvärdering del



Förnamn Efternamn

EC Utbildning

Kunskapskontroll, R-programmering

2024-04

.

Innehållsförteckning

[1 Teoretiska frågor 1](#_Toc165042339)

[2 Självutvärdering och grupparbetet 3](#_Toc165042340)

[2.1 Grupparbete 3](#_Toc165042341)

[2.2 Självutvärdering 4](#_Toc165042342)

# Teoretiska frågor

(1) Teoretiska frågor Besvara följande teoretiska 7 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.

QQ-plot är ett sätt att kolla om en distribution är normal-fördelad. Genom att sortera data n från lägst till högst och jämföra med en standardiserad normalfördelning som är delad in till n+1 delar. Alltså n jämförs då med den standardiserade normalfördelningen och exemplevis borde n\_1 vara nära den första quantilen av den standardiserad normalfördelningen om n är normalfördelad. Grafiskt sätt borde alla punkter lägga sig linjärt om och dem inte gör det kan de tyda på att vi har extrema värden i våra punkter vilket kan betyda att vi inte har en normalfördelad data

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?

Regressionsanalys och maskininlärning har sina gemener men skillnaden är att i Maskininlärning handlar om ”field of study that gives computers the ability to learn without beeing explicity programmed to”: Arthur Samuel 1956. Vad man menar med det är att man jobbar med att kunna lära en modell att göra prediktioner med data.

I regressionsanalys så jobbar man också med modeller men man utför också statistik inferens på variabler från datan för att kunna se relationer mellan variabler. Man kan utföra konfidens intervall, hypotestester, bedöma signifikansnivå osv.

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

Ett konfidensintervall av de predikterade värdena kommer att visa vilka värden som my kan anta med en oftast 95% sannolikhet givet att värdena är normalfördelade. Prediktions intervallet kommer göra exakt samma sak men kommer också innehålla epsilon alltså ett medelfel eftersom modellen inte perfekt predikterar nya observationer därför kommer prediktionsintervallet alltid att vara större än konfidensintervallet

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

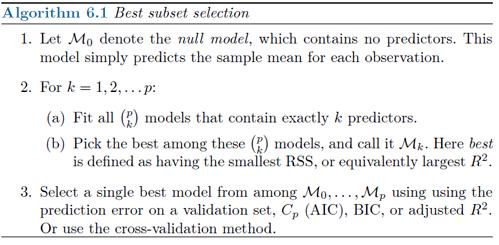
Betaparametrana tolkas som koefficienter som ändrar på den beroende variabeln givet att alla andra oberoende variabler är konstanta. Alltså hur förändras y när x\_n ändras.

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?

Antagandet bakom BIC, AIC, Cp osv antar att eftersom σhatt^2 är en o-bias skattning av σ^2 för epsilon så kan anta att Ex BIC är en o-bias skattning av modellens test error. Som konsekvens kommer BIC eller alla dessa mått anta ett lågt värde när modellens test error kommer va lågt

ekvation 8: notera att log() är den naturliga logaritmen.

6. Förklara algoritmen nedan för ”Best subset selection”



Best subset selection är en algoritm för att kunna ta fram de bästa predictors givet ett maxantal använda oberoende variabler nämnt K här. Man steg vis testar modeller med 0 till k variabler i sig och sedan jämför med varandra. Ex m\_0 har 0 p alltså bara ett intercept, m\_1 har ett p och då har modellen tagit den predictorn som ger bäst resultat, m\_2 har 2 p och då den kombinationen av predictors som ger bäst resultat. Detta görs om för alla modeller tills man har slutligen p+1 modeller att jämföra. Konstatera att denna algoritmen tränar och testar totalt sätt 2^p modeller som kan kräven väldigt mycket kraft vid högt antal p.

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

Om man bryter ner citatet kan man säga att ”All models are wrong” så syftar George på att ingen modell kan fånga komplexiteten från den riktiga världen och ge oss perfekta resultat utan modellen kan vara ge oss approximeringar.

”some are useful” så kan man argumentera för att han syftar på att modellerna ger oss tillräckligt med information för att vara användbara, ingen modell behöver vara perfekt för att vara användbar och det kan va viktigt att komma ihåg om man vill göra en robust modell.

# Självutvärdering och grupparbetet

## Grupparbete

1. Vem du har arbetat i grupp med?

I min grupp så bestod vi först av 6 pers och sedan 7

(mig själv, Aikaterini, Girlie, Jacob, Lidiia, Manna och sen NIl)

Nil fick hoppa in i vår grupp eftersom hon missa gruppindelningen men det medförde inga konstigheter, vi förklarade vad för data vi samlat in.

1. Hur har ni i gruppen arbetat tillsammans?

Vi började med att konstatera hur mycket data vi skulle samla in, en person hade dock jobbat lite i förväg och kollade just på Volvo v60 bilar. Personlig ville jag testa att ha fler bilmärken med gav in ganska snabbt eftersom det blir enklare att bara ha v60 bilar.

Lite diskussioner om vilken data vi skulle samla in, någon i grupp ansåg att man skulle behöva minst 100 observationer per variabel vilket lät helt absurt, i en snabb googling ser jag 4 sidor som säger att över 10 observationer räcker så jag konstaterar att det kanske inte behövs så många så i istället kan med lite fler variabler, tillslut kom vi fram till 5 variabler + pris och vi kom fram till att samlade in minst 50 observationer var vilket sluta med 352 observationer.

En annan diskussion som dock upp var hur vi anteckna datan. Vissa ville att vi skulle kategorisera numeriska data och även rangordna bränsletyp till 0,1,2 vilket man kan göra men jag argumentera att det är bäst att låta data vara som den är så kallat ”raw” och sedan kan varje individ transformera om datan om man vill det. Jag fick i alla fall igenom att vi skulle behålla den numeriska datan numerisk.

Själva insamlingen va uppdelad till modellår så varje individ tog hand om att samla in data från v60 bilar med respektive modellår vilket funkade perfekt.

Jag efter att ha samlat in mina 50 observationer för jag var snabbast gjorde ett snabbt f-test och konstaterade i gruppchatten att dem variablerna vi hade kommer nog funka bra när vi har alla observationer och folk i gruppen höll med.

1. Vad var bra i grupparbetet och vad kan utvecklas?

Jag tror personligen att som grupp kan man jobba mer på att diskutera och låta andra uttrycka sina åsikter om vad man vill göra eller kolla på. Det gick väldigt snabbt att bara välja x antal variabler och begränsa sig till v60. Självklart så kan det vara svårt att få utrycka sig och ibland så vill inte vissa individer ha en åsikt vilket är helt okej det med.

Det fanns situationer där någon fråga mig vad som kan va bra det kunde liksom inte komma fram i denna gruppen. Då skall jag även konstatera att jag har jobbat i bilbranschen och byggt dessa v60 modeller samt har någorlunda koll på vad som värdesätter just dessa bilar.

Viktigt att konstatera att datan dock blev jättebra, den var inte perfekt men det hade varit tråkigt och insåg att ha för mycket variabler kan nog ge mer problem än vad man har tid för att hantera. Vi var dock väldigt snabba på att dela upp arbetet att samla in data vilket gjorde att vi fick datan väldigt snabbt. Vi satte även en deadline på alla skulle samlat in sin del till senast den kommande måndagen vilket funkade bra.

1. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Nämnde att jag jobbat i bilbranschen så kanske skulle uttrycka det men istället lät jag andra ta ledning för å få deras rött hörde. Att ha grupparbeten och diskussion kan bli rätt ostrukturerat så man skulle kunna jobba med att ha någon typ av struktur så att alla kan få va delaktiga och på så sätt få ett ut ett bättre arbete.

1. Finns det något du hade gjort annorlunda? Vad i sådana fall?

Jag har även varit handledare och haft lite ansvar som lagledare så jag har några erfarenheter i att få lag att fungera och det är viktigt att låta alla få sina röster hörda så att dem kan känna sig involverade. Men jag kan ha varit för slö på den fronten och kanske själv skulle tagit en ledarroll. Men det är en väldigt svår fråga eftersom jag inte känner alla. Alla människor fungerar olika i olika situationer.

## Självutvärdering

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

Jag va lite lost i vad jag skulle göra i början och vad jag ens ville kolla på eller komma framtill men slutligen så kom det typ bara ihop och nu sitter jag här med en rapport på mer än 5000 ord vilket inte va tanken. Jag tänkte från början nästan begränsa mig till 2-3 tusen ord men det bara flög iväg vilket kanske är negativt.

Det fick lite en rödtråd när jag fick data om ny-registrerade bilar från SCB för då fanns det liksom ett syfte med att göra denna undersökning mer än att bara göra det för att göra.

Tycker också att man skulle kunna ha mer referenser men det blir så onödigt mycket redan så som det är

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

Tror jag har nått VG kriterierna så VG, men absolut inte IG

1. Något du vill lyfta fram till Antonio?

Skulle kunna tänka mig att få mera programmerings uppgifter, att kolla på källkod och leka med den är bra men man lär sig bäst när man försöker lösa ett problem tycker jag.