Projekt MA1487- Matematisk modellering

Statistiska analyser kan användas för att komma fram till rimliga slutsatser eller fatta beslut. En utmaning när man genomför statistiska analyser är att presentera dem på ett lättförstått och överskådligt sätt. I detta projekt ska kommer ni få bekanta er med denna konst.

Data

I projekt så ska man hitta en datamängd, genomföra statistiska analyser och presentera dessa. Data kan till exempel hämtas från:

- statistikmyndigheten SCBs databas. Ett verktyg för att hämta data via Python finns här https://github.com/kirajcg/pyscbwrapper.
- Väderdata från smhi:s API. Där man kan hämta väderdata från många olika väderstationer. Ett kort exempel (ej fullständigt) finner ni på kurshemsidan.
- En web-scraper där ni själva plockar ut data från en eller flera webbsidor, t.ex. med beautiful soup.

Valet av datamängd är upp till er, så länge datan uppfyller följande kriterier:

1. Minst 3 variabler, till exempel temperaturen vid 3 väderstationer.
2. Minst 30 datapunkter per variabel.
3. Data hämtas dynamiskt från ett API eller via en web-scraper.

Uppgifter

Inlämningen av projektet ska bestå av en skriftlig rapport där uppgift 1-7 besvaras med text och plottar, samt inlämning av all kod och data som använts i er analys. Uppgifterna löses i valfritt programmeringsspråk.

Tider för den muntliga redovisningen kommer göras tillgängliga på Canvas. **Rapporten behöver vara inne senast 2 hela arbetsdagar före presentation.** Detta för att säkerställa att alla hinner förbereda sin opponering.

Efter den muntliga redovisningen och opponeringen kommer läraren att ställa frågor kopplade till **rapporten** och **kursens innehåll**.

Uppgift 1: Beskriv data

Introducera den data som valts och beskriv vad den visar och varifrån den kommer. Cirka 250 ord (halv A4). Var tydliga med vad de olika variablerna beskriver och i vilken enhet de är i. Det kan vara en god idé att ha en mindre tabell med ett urval från datan för att lättare beskriva mätvärdena.

Det ska också finnas en visuell representation av hur datamängden ser ut, samt tillhörande figurtext med förklaringar till vad som visas och om det finns några konstigheter (till exempel outliers i datan). Visualiseringen görs med lämplig graf, t.ex. stapeldiagram, linjediagram, scatterplot, cirkeldiagram etc. **Obs! Glöm inte att ange enheter på axlarna!**

Uppgift 2: Beskrivande statistik

Gör en tabell innehållande beskrivande statistik av din data. Denna ska innehålla medelvärde, standardavvikelse, max- och min-värde samt korrelationen mellan variablerna. Korrelationen kan också med fördel visualiseras i form av en heatmap (i python Seaborn.heatmap(korrelation)).

Till dessa tabeller ska också en kortare text om vad dessa värden säger om er data och om det går att dra några slutsatser utifrån den.

Uppgift 3: Beskrivande plottar

Gör minst en graf till för att visuellt analysera er data. Det kan till exempel vara ett histogram som jämförs mot normalfördelningen eller ett lådagram för att vissa hur spridningen av data ser ut.

Uppgift 4: Linjär regression

Utför en linjärregression av minst en av variablerna och ett tillhörande 95% konfidensintervall. Rapportera variablerna a och b i sambandet $y = a + b \cdot x$ samt punktskattningens konfidensintervall av dessa. Visualisera detta i en graf med den linjära modellen, konfidensintervallet och originaldata i samma figur.

Uppgift 5: Transformerad data

Ibland passar inte den data man har till en linjär modell. Då kan det ibland gå att lösa genom att transformera data med exempelvis med en logaritmisk funktion. Prova minst en transformation av din data och skapa en ny regressionsanalys. Plotta sedan den nya modellen tillsammans med originaldata och jämför med den tidigare modellen. **Obs! Glöm inte att transformera tillbaka modellen och er data innan ni plottar dessa. Annars kan ni inte göra en tydlig jämförelse mellan de två modellerna.**

Uppgift 6: Residualanalys

Beräkna residualerna, $e = y - \hat{y}$, för de två modellerna och plotta dessa. Hur ser de ut? Plotta residualerna mot normalfördelningen (i Python t.ex. genom Seaborn.distplot eller scipy.stats.probplot). Kommentera dessa plottar utseende och beskriv vilka slutsatser vi kan dra utifrån dessa. Finns det några beroenden? Hur väl följer residualerna en normalfördelning?

Beräkna också deras varians och argumentera för vilken modell vi bör använda utifrån dina resultat.

Uppgift 7: Sammanfattning

Skriv en sammanfattning av din analys och vilka slutsatser du kommit fram till, max halv A4.

Uppgift 8: Muntlig presentation

Förbered en kort presentation av din data, analys och slutsatser. Denna presentation kommer hållas muntligt i mindre grupper efter nyår (vecka 1 eller 2). Presentationen ska hållas mellan 10-15 minuter och opponeras på av en annan student.

Uppgift 9: Opponering

Opponera på en annan students arbete. Dels skriftligt i form av en text på 150-250 ord och muntligt i samband med presentationen av den andra studentens arbete. Den skriftliga versionen ska sedan skickas till studenten som du opponerat på samt läraren.

Uppgift 10: Uppföljande frågor från Lärare

Var förberedd på specifika frågor på din rapport och generella frågor kopplade till ämnena som behandlats i kursen. **Observera att frågor från alla delar av kursen kan förekomma men med fokus på förståelse av samband mellan kursens ämnen.**