SVARARK for: Tellende prosjekt i ISTx1003 - høst 2022

|  |  |  |
| --- | --- | --- |
| **Oppgave 1 (50%)** | **Oppgavetekst** | **SVAR** |
| Q1a.1) | Hvor mange observasjoner har vi i datasettet? Matcher det med lengden på hver del av veien og den totale veilengden? | Det er gjort 3355 målinger, dette antallet målinger tilsvarer ikke den totale lengden, det er ca. 1200 meter uten målinger. |
| Q1a.2) | Hva tror du en negativ verdi av `sporing` betyr i praksis? Hvordan tolker vi betydningen av `sporing\_trafikk`? | Negativ sporing kan bety fartshumper eller telehiv. Sporing\_trafikk menes trafikken på det strekke som måles. |
| Q1a.3) | Vi manglet originalt 17 av 3355 målinger av veibredden. Hvordan løste vi det? | Dere løste det ved bruk av gjennomsnitt eller forventningsverdien. |
| Q1a.4) | Hva er det fulle navnet på asfalttypen vi har mest av? (skjelettasfalt, asfaltbetong eller asfaltgrusbetong) | Det er mest asfalt betong. |
| Q1a.5) | Hvorfor tror du veibredden er mye større i starten av datasettet? | Dette kan være fordi veien er nærmere Trondheim og det er mer trafikk på disse strekkene. |
| Q1a.6) | SVV vil at nye veier skal holde i 20-30 år. Basert på dataene fra 2020, er det realistisk? Hvorfor/hvorfor ikke? Veien skal repareres når spordybden blir for stor. Blir veien reparert ofte nok? Gi kort begrunnelse. | Nei, det går ikke. Det er for høy slitasje på veiene for at de kan holde i 20-30 år. Noen veier er allerede over 20 mm, noe som indikerer at veiene ikke blir kjapt nok reparert. |
| Q1b.1) | Skriv ned ligningen for den estimerte regresjonsmodellen og forklar de ulike elementene. |  |
| Q1b.2) | Hva er den estimerte verdien til skjæringspunktet (intercept) beta\_0^hat, og hvordan vil du tolke den? |  |
| Q1b.3) | Vi ser at for `spordybde` er `coef` lik 0.2665. Hva er formelen som er brukt for å regne ut denne verdien? Hvordan vil du forklare dette tallet til noen som ikke har hørt om lineær regresjon? |  |
| Q1b.4) | For `spordybde` har vi også tallene 0.239 og 0.294 i kolonnene `[0.025 0.975]`. Hva er disse to tallene og hvordan tolker du dem? |  |
| Q1b.5) | Videre står det for `spordybde` at `P>|t|` er 0.000. Hvilken hypotese er testet her? Hva er konklusjonen fra hypotesetesten hvis vi bruker signifikansnivå 0.05? Hvordan henger dette sammen med tallene 0.239 og 0.294 fra forrige punkt? |  |
| Q1b.6) | Hvilke modellantagelser gjør vi i en enkel lineær regresjon? |  |
| Q1b.7) | Hva er en predikert verdi og hva er et residual? Skriv også ned relevante formler. |  |
| Q1b.8) | Hvordan kan vi bruke predikert verdi og residual til å sjekke modellantagelsene? |  |
| Q1b.9) | Vi får oppgitt tallet `R-Squared` til å være 0.098 (skrives ofte som 9.8%). R2 har i enkel lineær regresjon en sammenheng med korrelasjonskoeffisienten, men det er også en annen definisjon som er relatert til sum av kvadrerte residualer. Hvilken formel er det? Forklar alle symboler. Hvordan vil du forklare R2 til noen som ikke har hørt om enkel lineær regresjon? |  |
| Q1b.10) | Studer plottet av predikert verdi mot residual. Hvordan skal et slikt plott se ut hvis modellantagelsene er oppfylt? Hvordan vil du evaluere plottet? |  |
| Q1b.11) | Studer QQ-plottet av residualene. Hvordan vil du evaluere plottet? |  |
| Q1b.12) | Vil du konkludere med at modellen passer godt? |  |
| Q1c.1) | Oppsummer kort hva du ser i plottene. Fokus skal være om du tror at det er noen sammenheng mellom `sporing\_trafikk` (som respons) og de tre forklaringsvariablene (`spordybde`, `veibredde` og `asfalttype`). Hvilken asflattype har høyest verdi for sporing\_trafikk? |  |
| Q1c.2) | Tilpass modellen ved å bruke koden for steg 3 og 4 over, med den nye formelen. Skriv ned ligningen for den estimerte regresjonsmodellen. |  |
| Q1c.3) | Hvor mange regresjonsparametere er estimert? Hva er betydningen av de ulike regresjonsparameterene? |  |
| Q1c.4) | Sammenlign den estimerte regresjonskoeffisienten for `spordybde` i denne modellen med samme koeffisient i den enkle lineære regresjonsmodellen. Har disse to samme tolkning? |  |
| Q1c.5) | Hva er predikert sporing per ti tusen biler for en veistrekning med asfaltbetong, 10 mm spordybde og 4 meter veibredde? (Regn ut for hånd ved å bruke relevante tall fra `resultat.summary()`.) |  |
| Q1c.6) | Forklaringsvariabelen `asfalt` er kategorisk og vi har brukt en såkalt dummy-variabelkoding, der `A1` (skjelettasfalt) er referansekategorien. Er effekten av de andre asfalttypene på sporing per tusen biler signifikant forskjellig fra effekten for asfaltbetong på nivå 0.05? Hvis vi sammenligner tre deler av veien med lik spordybde og veibredde, men ulik asfalttype, hva er gjennomsnittlig forskjell i sporing per tusen biler? Hvilken type asfalt ser ut til å gi den mest solide veien? |  |
| Q1c.7) | Hva er andel forklart variasjon? Ville du forventet at andelen forklart variasjon gikk opp da vi la til flere forklaringsvariabler enn bare `spordybde`? Hvis vi nå la til en forklaringsvariabel som forteller noe om dyrelivet langs veien, ville da R2 økt? |  |
| Q1c.8) | Basert på utskrifter og plott, vil du konkludere med at modelltilpasningen er god? |  |
| Q1c.9) | Utfør en ny multippel lineær regresjon (steg 2-5) med `sporing\_trafikk` som respons og `veibredde` og `spordybde` som forklaringsvariabler. Du må nå modifisere modellformelen ved å ta bort `asfalt`, og så kopiere inn kode for steg 2-5. Hvor mange regresjonsparametere er nå estimert? Hva er de signifikante forklaringsvariablene? |  |
| Q1c.10) | Er modelltilpasningen god? |  |
| Q1c.11) | Sammenlign `Adj. R-squared` (også kalt justert R2) for modellen med og uten `asfalt`. Hvis vi skal avgjøre om `asfalt` skal være med som forklaringsvariabel ved å bruke justert R2, hva blir konklusjonen? Begrunn svaret ditt. |  |
|  |  |  |
| **Oppgave 2 (30%)** | **Oppgavetekst** | **SVAR** |
| Q2a.1) | Hvorfor ønsker vi å dele dataene inn i trening-, validering- og test-sett? |  |
| Q2a.2) | Hva brukes hver av disse delene til i våre analyser? |  |
| Q2a.3) | Hvor stor andel av dataene er nå i hver av de tre settene? Ser de tre datasettene ut til å ha lik fordeling for de tre forklaringsvariablene og responsen? |  |
| Q2a.4) | Kommenter hva du ser i plottene og utskriften. |  |
| Q2a.5) | Hvilke av de tre variablene tror du vil være gode til å bruke til å predikere om det blir hjemmeseier? Begrunn svaret kort. |  |
| Q2b.1) | Hvilke forklaringsvariabler er signifikante i modellen på signifikansnivå 0.05? |  |
| Q2b.2) | Hvordan kan du tolke verdien av `exp(skudd\_paa\_maal\_diff)`? |  |
| Q2b.3) | Hva angir feilraten til modellen? Hvilket datasett er feilraten regnet ut fra? Er du fornøyd med verdien til feilraten? |  |
| Q2b.4) | Diskuter hva du ser (modell uten forseelse\_diff). |  |
| Q2b.5) | Som din beste modell for logistisk regresjon, vil du velge modellen med eller uten `forseelse\_diff` som kovariat? Begrunn svaret. |  |
| Q2c.1) | Forklar kort hva som er gjort i koden over, og hvilken verdi av k du vil velge. |  |
| Q2d.1) | Vil du foretrekke å bruke logistisk regresjon eller k-nærmeste-nabo-klassifikasjon på fotballkampdataene? |  |
| Q2d.2) | Oppsummer hva du har lært at kan være en god metode for å predikere om hjemmelaget vinner en fotballkamp. |  |
|  |  |  |
| **Oppgave 3 (20%)** | **Oppgavetekst** | **SVAR** |
| Q3a.1) | Hvilke 3 siffer har vi i datasettet? (Alle 3 sifrene er representert blant de første 10) |  |
| Q3a.2) | Hvor mange bilder har vi i datasettet? |  |
| Q3a.3) | Hvilket siffer ligner det 500. bildet i datasettet vårt på? Lag et bilde som viser dette sifferet. (Husk at Python begynner nummereringen med 0, og derfor refereres det 500. bildet til `[499]`) |  |
| Q3b.1) | Tegn sentroidene av de 3 klyngene fra $K$-gjennomsnitt modellen. Tilpass koden over for å plotte. Her kan du ta skjermbilde av sentroidene og lime inn i svararket. |  |
| Q3b.2) | Synes du at grupperingen i klynger er relevant og nyttig? Forklar. Maks 3 setninger. |  |
| Q3b.3) | Vi har valgt K = 3 for dette eksempelet fordi vi vil finne klynger som representerer de 3 sifferene. Men generelt er K vilkårlig. Kom opp med et forslag for hvordan man (generelt, ikke nødvendigvis her) best kan velge K. (Se her, for eksempel: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb). Beskriv i egne ord med maks 3 setninger. |  |
| Q3b.4) | Kjør analysen igjen med K = 2 og K = 4. Synes du de nye grupperingene er relevante? |  |
| Q3c.1) | Vurder dendrogrammet nedenfor. Synes du at den hierarkiske grupperingsalgoritmen har laget gode/meningfulle grupper av bildene? |  |
| Q3c.2) | I koden under har vi brukt gjeonnomsnittskobling (`method = 'average'`). Hvordan fungerer gjeonnomsnittskobling? Maks 2 setninger. |  |
| Q3c.3) | Velg en annen måte enn `method = 'average'` til å koble klyngene sammen (vi har lært om dette i undervisningen) og lag et nytt dendogram ved å tilpasse koden nedenfor. Kommenter resultatene. Ser det bedre/verre ut? |  |
| Q3d.1) | Hvis vi skulle brukt en metode for å predikere hvilket siffer fra et håndskrevet tall er, og ikke bare samle dem i klynge, hva ville du brukt? |  |