Prosjekt IST1003 Team 14

Eirik Steira, Stian Mogen, Nicolay Schiøll-Johansen

Oppgave 1: Regresjon

Q1.1:

a) *Skriv ned ligningen for den estimerte regresjonsmodellen. Forklar de ulike elementene.*

$$Y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim N(0, \sigma),\ i = 1, ..., n$$

Dette er likningen for en regresjonslinje, en rett linje som beskriver hvordan responsvariablen y endrer seg når forklaringsvariablen x endrer verdier. Denne linjen predikerer verdien av y for en gitt verdi av x og er gitt ved formelen over.

B0 er skjæringspunktet i y-aksen, den verdien y tar for x=0.

ei tar hensyn til variabiliteten i datasettet, altså umålte ting. Dette er det vi ikke har forklart ved forklaringsvariablen gjennom regresjonslinjen, kalt residualer. En residual for hver observasjon er gitt ved formelen ei = yi - (B0+B1xi). ei antas uavhengige og normalfordelt med forventningsverdi 0 og standardavvik sigma.

B1 er stigningstallet til linjen, mengden y endrer seg når x endrer seg med en enhet.

B1xi forklarer “avvik” fra gjennomsnittet. Hvis to personer, A og B, har alt likt uten én kovariant, si A har en enhet høyere enn B, så vil i gjennomsnitt person A ha en prognosescore som er én verdi av B1xi høyere eller lavere enn B.

b) *Hvordan vil du tolke den estimerte verdien til skjæringpunktet (Intercept) beta0.hatt?*

Den estimerte y-verdien for en idrettsutøver med 0 for alle kovariater. B1xi = 0 for alle xi.

beta0.hatt fungerer som konstantleddet til den estimerte linjen for antall blodceller basert på høyde, altså y der xi = 0. Verdien for beta0.hatt er 1.0669. Vi må summere denne konstanten med beta1.hatt * xi for å finne y-veriden i den estimerte linjen for hver enkel xi.

Q1.2:

a) *Vi ser at for 'Hoeyde' er 'coef' lik 0.0199. Hva er formelen som er brukt for å regne ut denne verdien? Hvordan vil du forklare dette tallet til en medstudent som ikke har hørt om enkel lineær regresjon?*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Formelen for å regne ut dette tallet er beta1.hatt = (Sum(xi-x.gj-snitt) * (Yi-Y.gj-snitt)) / Sum(xi-x.gj-snitt). Dette er gjennomsnittlig avvik fra gjennomsnittet til x ganger gjennomsnittlig avvik fra regresjonslinjen delt på gjennomsnittlig avvik fra gjennomsnittet til x.

beta1.hatt er stigningstallet til den estimerte regresjonslinjen. Enkelt forklart så vil dette være estimatet for økning av antall blodceller lineært med økningen av centimeter høyde. Hver centimeter økning gir 0.0199 økning i blodceller-

b) *For 'Hoeyde' er det også gitt de to tallene 0.013 og 0.027 under kolonnene "[0.025 0.975]". Hva er disse to tallene og hvordan tolker du tallene?*

Disse tallene gir oss vårt 95% konfidensintervall. Intervallet strekker seg altså fra 0.013 til 0.027 og gir oss en vurdering på feilmarginen på målingene våre. Den virkelige verdien for b1.hatt.hoyde ligger med ganske stor sikkerhet innenfor disse to verdiene. “Hoeyde” gir oss også coef = 0.0199, som er den beste gjetningen vi har på hvor B_1 ligger. 0.0199 ligger innenfor 95% konfidensintervallet vårt, så det lover godt.

c) *Videre står det for 'Hoeyde' at 'P>|t|' er 0.000. Hvilken hypotese har man testet her? Hva er konklusjonen fra hypotesetesten hvis vi bruker signifikansnivå 0.05? Hvordan henger dette sammen tallene 0.013 og 0.027 fra forrige punkt?*

Vi tester nullhypotesen der Hnull : beta1 = 0, eller H1 : beta1 != 0. I praksis tester vi om regresjonslinjen er flat eller skrå. Men signifikansnivå 0.05 kan vi bruke z0.025 = 1.96, og siden t = 5.669 > 1.96, kan vi altså forkaste Hnull til fordel for H1. Linjen er altså skrå, siden absoluttverdien til testobservatoren er større enn den kritiske verdien. Vi ser at vi har 95% konfidensintervall for at stigningstallet er mellom 0.013 og 0.027, som passer med konklusjonen vår om at linjen er skrå (altså stigningstall ulik 0).

Q1.3:

a) *Hvilke modellantagelser gjør vi i en enkel lineær regresjon?*

1. Det er en lineær sammenheng mellom forklaringsvariabelen og responsen, x og y.

2. Feilleddene har konstant varians - for alle verdier av forklaringsvariabelen.

3. Feilleddene er normalfordelte

4. Observasjonsparene (og da også feilleddene) er uavhengige.

b) *Hva er en predikert verdi og hva er et residual (formler)?*

En predikert verdi er en prediksjon for hvilken verdi responsen kan få for en eller flere gitte verdier i forklaringsmodellen. Dette er for å forstå sammenhengen mellom en forklaringsvariabel og en respons og ved å bruke predikerte verdier kan vi bruke den estimerte modellen som en prediksjonsmodell.

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

Vi kan bruke den estimerte regresjonslinjen til å predikere en verdi for responsen for den nye observasjonen. Dette gjøres ved å bruke formelen y0.hatt = beta0.hatt + beta1.hatt.x0. Da kan vi også regne ut residualer .

Residualer er vår differensen mellom observert og predikert verdi av responsen. Det er vårt beste gjett, eller prediksjon, for feilleddet. Et residual er gitt ved formelen ei.hatt = yi - yi.hatt.

c) *Hvordan kan vi bruke predikert verdi og residual til å sjekke modellantagelsene?*

For å sjekke modellantagelsene kan vi plotte residualer mot predikerte verdier og lage et QQ-plott av residualene der vi sammenligner kvantiler for den empiriske fordelingen til residualene med kvantiler i normalfordelingen.

Dersom vi finner en trend i residualene vil det bety at regresjonsmodellen ikke har fått all informasjon fra forklaringsvariabelen x. Dette tester punkt 1 i oppgaven over. Samtidig vil variansen til feilleddet ikke nødvendigvis være konstant hvis bredden på området for residualer ikke er konstant, noe som tester punkt 2. Til slutt, hvis residualene ligger på en rett linje i QQ-plottet vil de være normalfordelte, noe som tester punkt 3.

d) *Vi får også oppgitt tallet "R-Squared" til å være 0.238 (ofte også skrevet som 23.8%). R^2 har i enkel lineær regresjon en sammenheng med korrelasjonskoeffienten, men det er en annen definisjon som er relatert til sum av kvadrerte residualer. Hvilken formel er det? Forklar alle symboler. Hvordan vil du forklare tallet til en medelev som ikke har hørt om enkel lineær regresjon?*

$$R^2 = \frac{SST - SSE}{SST} \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SST: Total sum of squares. Summen av alle kvadrerte avvik av yi fra gjennomsnittet av y-ene. Denne summen av kvadrerte avvik fra gjennomsnittet er den samme som den vi

bruker når vi estimerer variansen til en stokastisk variabel y (når vi ikke bryr oss om noen sammenheng med annen variabel x).

SSE: Error sum of squares. Summen av alle kvadrerte avvik fra yiHatt, der yiHatt er posisjonen til observasjonen i på den estimerte regresjonslinjen. Altså sum av avvik fra den estimerte regresjonslinjen. Denne summen brukes når vi estimerer variansen til en variabel y betinget på x.

R-squared er med andre ord andelen varians i Y som forklares av den lineære regresjonsmodellen.

Q1.4:

a) *Studer plottet av predikert verdi mot residual. Hvordan skal et slikt plott se ut hvis modellantagelsene er oppfylt? Hvordan vil du evaluere plottet?*

Dersom modellantagelsene er oppfylt skal gjennomsnittet av residualene i et slikt plott være lik null og vi skal ha omtrendt samme variabilitet hele tiden, altså feilleddene er normalfordelt med konstant varians.

I dette plottet ser gjennomsnittet ut til å ligge rundt null og har omtrent samme variabilitet hele tiden. Det konkluderes derfor med at plottet passer greit til modellen.

b) *Studer QQ-plottet av residualene. Hvordan vil du evaluere plottet?*

Punktene ligger i dette tilfellet nære den røde linjen, slik at den følges ganske tett. Vi kan derfor si at antagelsen om normalfordelte feilledd ser ut til å stemme, vi har ut fra dette ikke grunnlag til å si at modelantagelsen ikke stemmer.

c) *Vil du konkludere med at modellen passer godt?*

Med bakgrunn i svarene på oppgaven over kan vi si at antagelsen om at variansen er den samme for alle kovariater og at residualene er normalfordelte, kan vi si at modellen passer godt.Plottet viser ingen trend og omtrent konstant bredde.

Q1.5:

Oppsummer kort hva du ser i plottene. Fokus skal være om du tror at det er noen sammenheng mellom Blodceller (som respons) og de fire mulige forklaringsvariablene (Hoeyde, Vekt, Kjoenn og Sport). Hvilket Kjoenn har generelt høyest verdi for Blodceller?

Ved å se på de første plottene, ser det ut som det er relativt liten sammenheng mellom blodceller, og høyde/vekt. Fra tidligere i oppgave Q1.2.C har vi allerede hypotesetestet sammenhengen mellom høyde og blodceller, og vi ser at det er en viss sammenheng, da vi ser at b.hatt1 er ulik null. Det ser ut som at det er en tydeligere sammenheng mellom høyde og blodceller, sammenlignet med vekt og blodceller, hvor sammenhengen ser mer ut til å være tilfeldig fordelt.

Ser vi på plottene som sammenligner kjønn, virker det å være ganske tydelig at menn generelt sett har flere blodceller enn kvinner. (Her antar vi at kjønn 0 er men og kjønn 1 er kvinne).

Basketball ligger noe lavere i antall blodceller enn de andre idretten, men det er vanskelig å avgjøre om det er tilfeldig eller ikke. Det virker også litt merkelig, da vi har konkludert tidligere at høyde kan indikere flere blodceller, og at basketballspillere ofte er høyere enn andre utøvere.

Q1.6:

a) *Skriv ned ligningen for den estimerte regresjonsmodellen. Hvor mange regresjonsparametre er estimert?*

$$Y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim N(0, \sigma),\ for\ i = 1, ..., 8$$

Det er 8 regresjonsparametre som er estimert; Intercept, Kjønn, 4 sporter (roing, svømming, tennis, turn), vekt og høyde.

For denne modellen vil det i praksis være:

$$Y_i = B_0 + B_1 x_{roing} + B_2 x_{svømming} + B_3 x_{tennis} + B_4 x_{turn} + B_5 x_{kjønn} + B_6 x_{høyde} + B_7 x_{vekt}$$

b) *Sammenlign den estimerte regresjonskoeffisienten for Hoeyde i denne modellen mot den estimerte regresjonskoeffisienten for Hoeyde i den enkle lineære regresjonen. Har disse to samme fortolkning?*

Ny modell: 0.0126, gammel modell: 0.0199

Hvis vi sammenligner to personer som ellers er like, så vil en person med en enhet mer ha en responsverdi som er beta.Hoeyde større enn den første. Da har den andre ??. en differanse på 1 cm i høyden på en person vil i gjennomsnitt gi en forskjell i blodceller på 0.0073.

I den nye modellen har høyde mindre å si for blodceller.

c) *Hvis vi sammenligner en mann og en kvinne som begge er like høye, veier like mye og begge holder på med samme idrett, hva er gjennomsnittlig forskjell i antall blodceller mellom dem?*

Gjennomsnittlig forskjell i antall blodceller mellom en mann og en kvinne som ellers har like kovarianter vil tilsvare forskjellen i den estimerte regresjonskoeffisienten for kjønn, altså 0.7131 lavere enn hos menn i dette tilfellet.

d) *Hva er predikert antall blodceller for en mann som holder på med roing, er 180 høy og veier 75 kg?*

(Regn for hånd ved å putte inn tall fra resultat.summary().)

$$\widehat{Y} = 3.6570 + 1 * 0.2165 + 180 * 0.0126 * 75 * (-0.135) = 5,129\ blodceller$$

Q1.7:

a) *Forklaringsvariablen 'Sport' er kategorisk og vi har brukt en såkalt dummy-variabelkoding, der 'Basketball' er referansekategorien. Er effekten av de andre sportskategoriene på 'Blodceller' signifikant forskjellig fra effekten for 'Basketball' (på nivå 0.05)?*

For å finne ut om effekten av de andre sportskategoriene på 'blodceller' er signifikant forskjellig fra effektekten for 'Basketball' på nivå 0.05 må vi se på p-verdien i utskriften. Dette er oppfylt for p > |t| er større enn 0.05.

Testing av koeffisienten for roing (0.003):

H0: I en modell der kjønn, vekt og høyde er med, er det ikke forskjell i antall blodceller for basketball og roing: $\beta 2=0$

H1: I en modell der kjønn, vekt og høyde er med, er det forskjell i antall blodceller for basketball og roing: $\beta 2\neq 0$

Nullhypotesten forkastes og roing er dermed viktig for å forklare antall blodceller.

Testing av koeffisienten for svømming (0.238):

H0: I en modell der kjønn, vekt og høyde er med, er det ikke forskjell i antall blodceller for basketball og svømming: $\beta 2=0$

H1: I en modell der kjønn, vekt og høyde er med, er det forskjell i antall blodceller for basketball og svømming: $\beta 2\neq 0$

Nullhypotesten forkastes ikke og svømming er dermed ikke viktig for å forklare antall blodceller.

Testing av koeffisienten for tennis (0.008):

H0: I en modell der kjønn, vekt og høyde er med, er det ikke forskjell i antall blodceller for basketball og tennis: $\beta 2=0$

H1: I en modell der kjønn, vekt og høyde er med, er det forskjell i antall blodceller for basketball og tennis: $\beta 2\neq 0$

Nullhypotesten forkastes og tennis er dermed viktig for å forklare antall blodceller.

Testing av koeffisienten for turn (0.259):

H0: I en modell der kjønn, vekt og høyde er med, er det ikke forskjell i antall blodceller for basketball og turn: $\beta 2=0$

H1: I en modell der kjønn, vekt og høyde er med, er det forskjell i antall blodceller for basketball og turn: $\beta 2\neq 0$

Nullhypotesten forkastes ikke og turn er dermed ikke viktig for å forklare antall blodceller.

Vi konkluderer derfor med at svømming og turn har signifikant forskjellig effekt, mens roing og tennis ikke har det.

b) *Hva er andel forklart variasjon? Ville du forventet at andelen forklart variasjon gikk opp da vi la til flere forklaringsvariabler enn Hoeyde? Hvis vi nå la til en forklaringsvariabel som var IQ til idrettsutøveren, ville da R^2 økt?*

Andel forklart variasjon er gitt ved R^2=0,711.

Hvis vi legger til flere forklaringsvariabler kan vi forvente at den forholder seg på samme nivå eller øker. Med utgangspunkt i dette og oppgaven over kunne vi forventet at denne gikk opp da vi la til flere forklaringsvariabler enn Hoyde. Dette fordi vi har flere variabler som forklarer dataen.

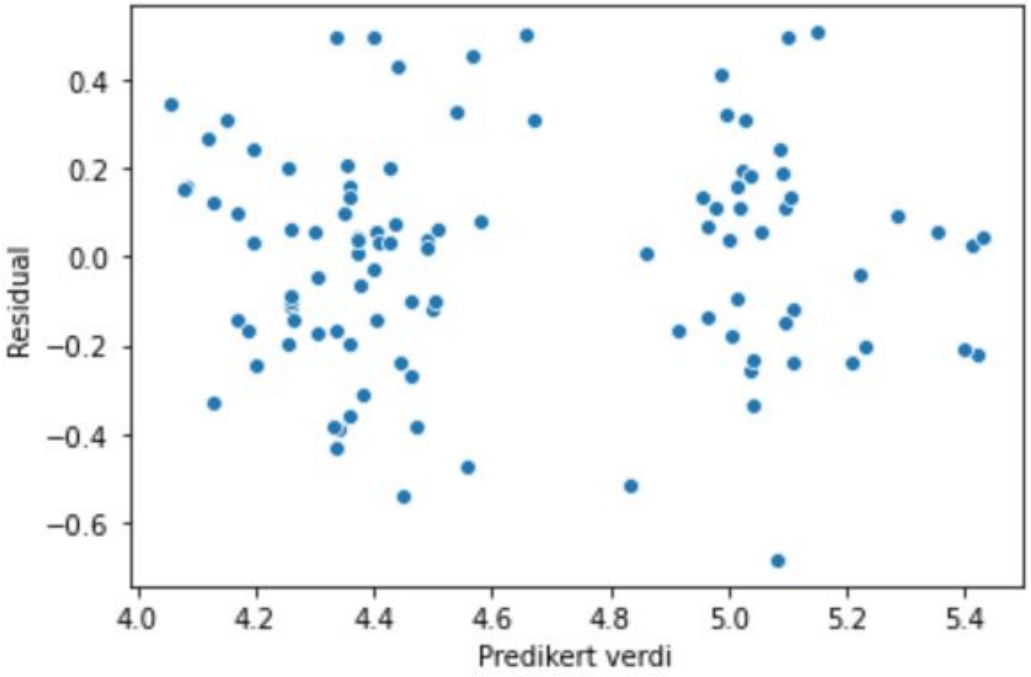
Dersom vi hadde lagt til en ny forklaringsvariabel, IQ, kunne vi derfor forventet at R^2 forholdt seg uendret eller økte. Siden IQ kan påstås å ikke ha en direkte tilknytning til antall blodceller, kan vi si at å inkludere denne forklaringsvariabelen vil innføre såkalt tilfeldig støy. Likevel vil R^2 kunne øke dersom det inneholder trender i dataen. Vi må derfor se på den justerte R^2 for å kunne sammenlikne to modeller med forskjellig antall forklaringsvariabler. I modellen vår er denne verdien Adj. R^2 = 0.690

c) *Basert på utskrifter og plott. Vil du konkludere med at modelltilpasningen er god?*

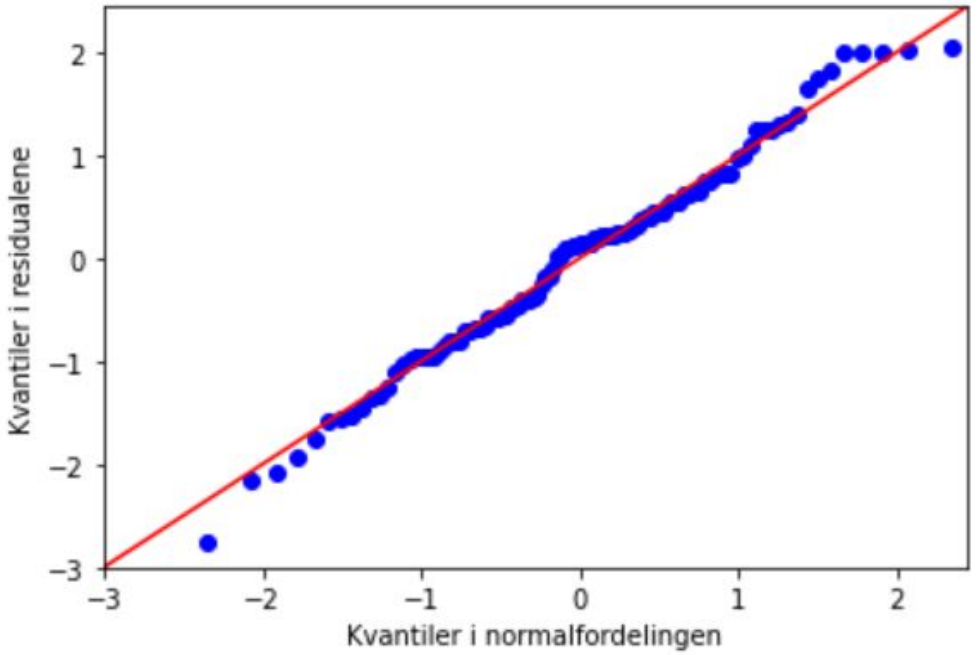
For å vurdere hvor god modelltilpasningen er må vi se på hvor stor andel av variasjonen i dataene vi har forklart med regresjonsmodellen, R^2 . Vi har at R^2 = 0.711 slik at vi har forklart 71,1% av variabiliteten i dataene. For å utelukke forklaringsvariabler som ikke er signifikante, bruker vi Adj. R^2 = 0.690, eller 69%.

regelen er at vi velger den modellen som har høyest justert-R^2. I denne modellen er denne verdien høyere enn i den forrige som var lik 23% og vi kan derfor konkludere med at denne modelltilpasningen vi har er bedre.

Når modellen passer vil plottet vise ingen trend og konstant bredde. I plottet over ser vi derimot av vi har to trender, selv om verdienes ser ut til å ha konstant bredde. Disse finner vi mellom 4.2-4.6 og 4.8-5.2.



Vi ser også at dette stemmer med antagelsen om at variansen er den samme for alle kovariater og tror derfor at feilleddene har konstant varians.



Vi ser at observasjonene ikke ligger på en rett linje og kan derfor konkludere med at residualene ikke er normalfordelte. Derfor tror vi heller ikke at feilleddene er normalfordelte.

Med bakgrunn i observasjonene over kan vi konkludere med at modelltilpasningen er noenlunde god, men kunne vært bedre. Vi bør jobbe for å finne en bedre tilpasning.

Q1.8:

a) *Hvor mange regresjonsparametere er nå estimert? Hva er signifikante forklaringsvariabler?*

$$\widehat{Y}_i = \widehat{B}_0 + \widehat{B}_1 * x_{kjønn} + \widehat{B}_2 * x_{\ høyde} + \widehat{B}_3 * x_{vekt} + e_i$$

$$\widehat{Y}_i = 4.3316 - 0.7632 * x_{kjønn} + 0.0089 * x_{høyde} - 0.0114 * x_{vekt}$$

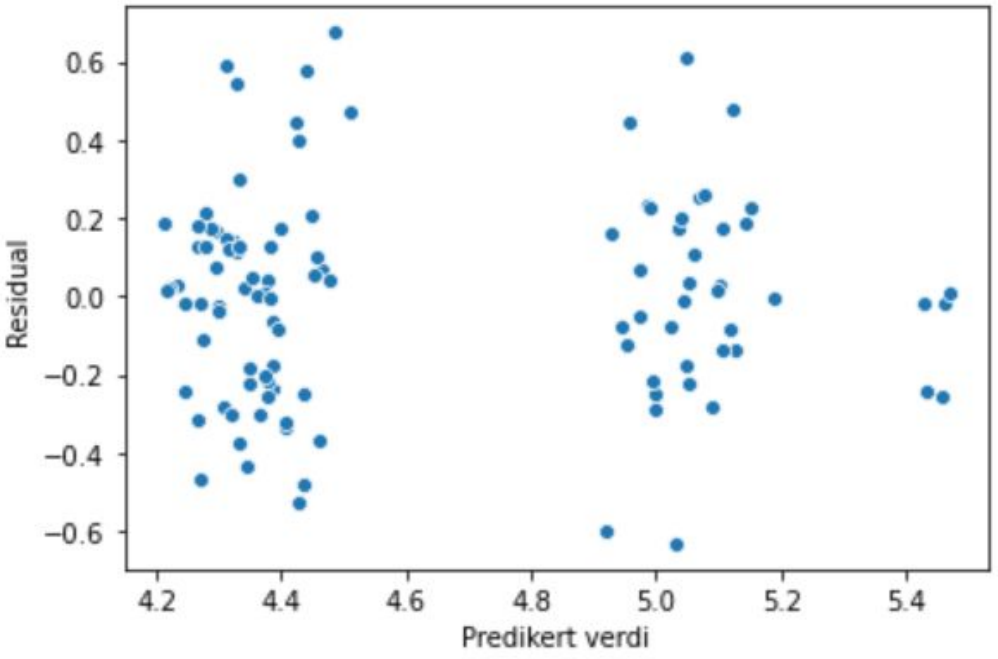
Vi har at 4 regresjonsparametre er estimert.

For å finne ut hvilke av disse som er signifikante må vi se på p-verdien men et signifikansnivå 0.05. Ingen av forklaringsvariablene er signifikant forskjellig, som vil si at kjønn, høyde og vekt alle er signifikante forklaringsvariabler.

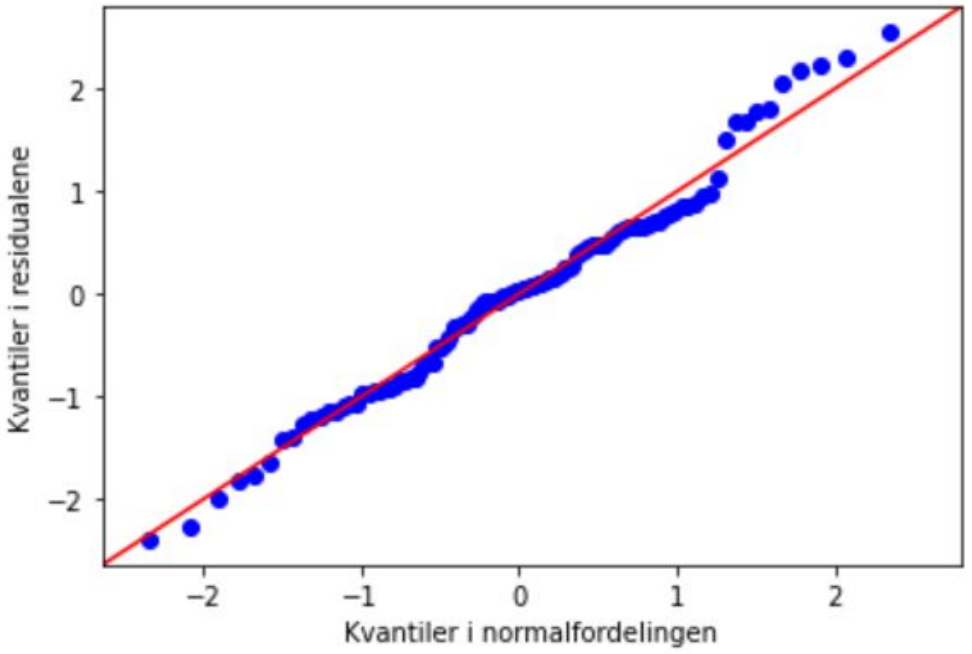
b) *Er modelltilpasningen god?*

Nye justert R^2 er 0.661, eller 61.1%, som er lavere enn den forrige modelltilpasningen, men fortsatt relativt høy.

Videre ser vi har vi har to enda tydeligere trender i residualplottene. Likevel ser de ut til å ha relativt konstant varians og vi tror derfor at feilleddene også har konstant varians.



I dette QQ-plottet kommer vi frem til samme konklusjon som i oppgave 1.7 c), at residualene ikke er normalfordelte.



Vi kan derfor konkludere med at modelltilpasningen ikke er like god og bør forbedres.

c) *Sammenlign Adj. R-squared for modellen med og uten 'Sport'. Hvis vi skal avgjøre om 'Sport' skal være med som forklaringsvariabel ved å bruke Adj. R-squared, hva vil du da konkludere med? Begrunn valget ditt.*

I modellen med 'Sport' har vi at Adj. $R^2 = 0.711$, mens den uten er $R^2 = 0.661$. Vi ser at modellen forklarer en større andel av dataene når 'Sport' er med. Vi ønsker altså at denne verdien skal så høy som mulig. I tillegg tar verdien hensyn til usignifikante forklaringsvariabler.

Vi konkluderer derfor med at kategorien 'Sport' bør være med i modellen, siden den resulterer i høyeste verdi for Adj. R^2.

Oppgave 2: Klassifikasjon

Q2.1:

a) *Hvorfor ønsker vi å dele dataene inn i trening, validering og test-sett?*

Når vi klassifiserer observasjoner har vi fokus på prediksjon og ønsker å lage en regel som også generaliserer til nye data. Vi passer på at regelen oppfyller dette og ikke passer for godt til våre egne data ved å dele dem inn i trening, validering og test-sett. Dette også fordi

ønsker vi en sannsynlighet for at observasjonene tilhører hver av klassene - både regelen og sannsynligheten skal kunne generaliseres til nye data og brukes i fremtiden.

***b)** Hva brukes hver av disse delene til i våre analyser?*

Treningssettet bruker vi til å lage en klassifikasjonsregel og består som regel av ca 60% av dataene våre - vi bruker det til å estimere modellparametrene slik at modellen vår tilpasses datene våre.

Valideringssettet bruker vi til å velge modell- og hyperparametre. Det kan altså brukes til å bestemme hvor fleksibel vi vil ha "grensen" - en metode for å bestemme klassen til en ny observasjon basert på de nærmeste naboene basert på disse hyperparametrene. Dette datasettet er som regel på ca 20% av dataene våre.

Testsettet bruker vi helt til slutt for å evaluere regelen på fremtidige data - hvor god regelen er for ny og fremmed data. Vi ser spesielt etter hvordan metoden fungerer på nye data og hvordan metoden fungerer på treningssettet er ikke like relevant, da vi kan oppnå såkalt "overtilpasning" der vi tilpasser metoden veldig bra til treningsdataene, men ikke nødvendigvis til nye data.

***c)** 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?*

Etter inndelingen består treningssettet av 60%, validerngssettet av ca 20% og testsettet av 20% av dataene. De tre datasettene ser ut til å veldig lik fordeling mellom de tre forklaringsvariablene og responsen. En forskjell som kan være verdt å merke seg er at for y i treningsettet under 50%, ser vi at vi får verdien 1.0, mens i de to andre settene er denne verdien 0.5.

Q2.2:

***a)** Kommenter hva du ser i plottene og utskriften.*

Hvis vi ser på normalfordelingen for acediff og upressetdiff ser vi en korrelasjon mellom antall ess og antall seiere og antall upressede feil og antall seiere. Flere ess for en spiller gir som regel flere seiere og motsatt for upressede feil. Dobbeldiff gir nokså lik respons og ser ut til å ikke like stor betydning.

Det plottet som kanskje gir den tydeligste korrelasjonen er den som er mellom upressede feil og dobbeltfeil, hvor økningen er lineær mellom de forskjellige typene med feil. Dette indikerer at dersom en spiller gjør en type feil, vil spilleren ha en tendens til å også gjøre andre feil. Det ser altså ut som at forklaringsvariablene er avhengig av hverandre.

Vi ser også at medianen for forklaringsvariablene i spesielt ess og upressede feil er ulike for seier (1) og tap (0) i en match. Med bakgrunn i dette kan vi forhåpentligvis lage en god klassifikasjonsregel basert på forklaringsvariablene.

***b)** Hvilke av de tre variablene tror du vil være gode til å bruke til å predikere hvem som vant matchen? Begrunn svaret.*

Ved å se på plottet kan vi gjøre en antagelse om at acediff og upressetdiff er de to variablene som gir den beste indikasjonen på hvilken spiller som vinner.

Høyere antall ace gir større sannsynlighet for å vinne en match, beskrevet en positiv korrelasjon på 0,32 .

Økning i antall upressede feil gir motsatt effekt, men en negativ korrelasjon på 0.38.

Dobbeltfeil ser også ut til å ha en liten negativ effekt på antall seiere, men ikke like mye som upressede feil, med en negativ korrelasjon på 0.15.

Alle variablene kan derfor bidra til å predikere resultatet, men ingen av dem vil alene være spesielt gode til å gjøre dette alene. Acediff og upressetdiff har den sterkeste lineære sammenhengen av disse variablene, med henholdsvis 0.32 og -0.38, og disse vil derfor være best egnet til å predikere resultatet av tennismatchen.

Q2.3:

a) *Hvilke forklaringsvariabler er signifikante i modellen på signifikansnivå 0.05?*

Acediff og upressetdiff er signifikante i modellen på signifikansnivå på 0.05

Dette leser vi av tabellen da acediff har p-verdi 0.00 og |z| = 8.067, som gir oss p < |z| og p < 0.05.

Samme gjelder for upressetdiff med p-verdi 0.00 og |z| = |-8.498| = 8.498, som gir oss p < |z| og p < 0.05.

Dobbeldiff på den andre siden er ikke signifikant i modellen, da vi kan se at p-verdi 0.909 > |z| = 0.115.

b) *Hvordan kan du tolke verdien av exp(upressetdiff)?*

For logistisk regresjon så vil vi ikke oppleve samme endring for responsen når vi har en endring i en forklaringsvariabel, men endringene i suksessannsynligheten vil være avhengig av hva tilhørende verdi for observasjonen xi er. Når x øker med 1 så vil oddsen multipliseres med exp(B1), her exp(upressetdiff). Vi kan se på spesialtilfellet vårt:

B1til upressetdiff = -0.0900, Siden B1 er mindre enn 0, vil oddsen minke.

exp(upressetdiff) altså, e ^ B1 = 0.913963. Dette tallet multipliseres som nevnt med oddsen når x øker med 1. Oddsen ganges derfor med 0.913963.

c) *Hva angir feilraten til modellen? Hvilket datasett er feilraten regnet ut fra? Er du fornøyd med verdien til feilraten?*

Feilraten beskrives som andel feilklassifiserte observasjoner, og vi kan bruke valideringssettet for å finne modellen vi skal bruke. Modellformen her bruker forklaringsvariablene acediff, dobbeldiff og upressetdiff. For å regne feilraten tar vi antall feil delt på antall observasjoner. I dette tilfellet så får vi en feilrate på cirka 0.215, dette er en

ganske høy prosent som indikerer at vi kanskje burde samle inn flere data. Det kan likevel være verdt å nevne at det er vanskelig å med sikkerhet forutsi en tenniskamp og sport generelt, sånn sett er ikke en feilrate på 0.215 helt urimelig i dette tilfelle.

Q2.4:

a) *Diskuter hva du ser.*

```
Optimization terminated successfully.
         Current function value: 0.515312
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  471
Model:                          Logit   Df Residuals:                      467
Method:                           MLE   Df Model:                            3
Date:                Wed, 11 Nov 2020   Pseudo R-squ.:                 0.2565
Time:                        13:15:21   Log-Likelihood:                -242.71
converged:                       True   LL-Null:                       -326.46
Covariance Type:            nonrobust   LLR p-value:                 4.406e-36
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -0.0732      0.111     -0.657      0.511      -0.292       0.145
acediff         0.1895      0.023      8.067      0.000       0.143       0.236
dobbeldiff      0.0041      0.036      0.115      0.909      -0.066       0.074
upressetdiff   -0.0900      0.011     -8.498      0.000      -0.111      -0.069
==============================================================================
FLERE utregninger:
exp(beta):  Intercept       0.929390
acediff         1.208629
dobbeldiff      1.004114
upressetdiff    0.913963
dtype: float64
Feilrate: 0.21518987341772156
```

```
Optimization terminated successfully.
         Current function value: 0.515326
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  471
Model:                          Logit   Df Residuals:                      468
Method:                           MLE   Df Model:                            2
Date:                Wed, 11 Nov 2020   Pseudo R-squ.:                 0.2565
Time:                        13:15:21   Log-Likelihood:                -242.72
converged:                       True   LL-Null:                       -326.46
Covariance Type:            nonrobust   LLR p-value:                 4.270e-37
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -0.0736      0.111     -0.661      0.508      -0.292       0.145
acediff         0.1895      0.023      8.069      0.000       0.143       0.236
upressetdiff   -0.0895      0.010     -9.012      0.000      -0.109      -0.070
==============================================================================
FLERE utregninger:
exp(beta):  Intercept       0.929011
acediff         1.208645
upressetdiff    0.914343
dtype: float64
Feilrate: 0.21518987341772156
```

Vi ser her at dersom vi fjerner dobbeldiff, så vil ikke det ha noen betydning på feilraten vi får på valideringssettet. Denne er fortsatt tilnærmet lik 0.215, dette er en indikasjon på at dobbeldiff ikke forbedrer modellen vår.

b) *Som din beste modell for logistisk regresjon vil du velge modellen med eller uten dobbeldiff som kovariat? Begrunn svaret.*

Her ville vi valgt en modell uten dobbeldiff som kovariat. Feilraten endrer seg som nevnt ikke når vi fjerner den fra modellen, utregning viste oss også tidligere i oppgaven at den ikke var signifikant til modellen.

Q2.5:

Forklar kort hva som er gjort i koden over, og hvilken verdi av k du vil velge.

Her er feilraten til k-nærmeste nabo-klassifikasjon for ulike verdier av k (antall naboer) plottet for valideringssettet. Den skriver også ut første verdien for k med lavest feilrate. Vi går kun gjennom oddetallsverdier for k, slik at vi unngår like naboer.

Av tallene vi får oppgitt er verdi 47 av k det beste alternativet, ettersom dette er den største oddetalls verdien til k med den laveste feilraten (0.19620253). Om vi derimot utvider søket vårt litt, finner vi at k verdi 59 med samme feilrate er et bedre alternativ. Vi velger en stor verdi av k, siden vi ønsker en modell som ikke er alt for fleksibel.

Q2.6:

a) *Vil du foretrekke å bruke logistisk regresjon eller k-nærmeste-nabo-klassifikasjon på tennisdataene?*

Vi ser at logistisk regresjon har noe lavere feilrate 0.196 sammenlignet med den vi får ved k-nærmeste nabo 0.215. Vi ville av den grunn valgt k-nærmest-nabo, de er likevel ganske like, så vi mener at begge vil være gode å bruke på tennisdataene.

Vi kan også kommentere at feilraten til logistisk regresjon i testsettet er lik feilraten til k-nærmeste-nabo i valideringssettet. Dette er tilfeldig, og kommer som konsekvens av at vi har et ganske lite datasett.

b) *Oppsummer hva du har lært at kan være en god metode for å predikere hvem som vinner en tennismatch.*

Vi har sett på forskjellige måter å predikere hvem som vinner en tennismatch, både med tanke på modellen vi velger, samt hvilke forklaringsvariabler vi skal bruke.

Til å begynne med kan vi si at førsteinntrykket vi fikk ved å enkelt observere diagrammene stemte ganske godt med hva vi til slutt regnet oss frem til. Hypotesen vår var umiddelbart at dobbeldiff i mindre grad enn de to andre klassifiseringsvariablene ville hjelpe oss å predikere vinneren. Etter å ha regnet ut hvilke som var signifikante, samt sammenlignet feilrate, konkluderte vi til slutt med at dobbeldiff ikke var relevant for modellen vi ønsket. Derfor endte vi opp med en modell som kun bruker acediff og upressetdiff.

Videre ser vi at selv om variablene gir oss en viss indikasjon, er den ikke i nærheten av å kunne predikere utfallet av en tennismatch med 0 feilrate. Dette gir mening, da idrett generelt sett er vanskelig å forutsi, og feilraten er betraktelig bedre enn dersom vi bare gjettet vilkårlig. Om vi bruker logistisk regresjon vil vi få en feilrate på 0.196 for test-settet, noe som vi kan si oss ganske bra fornøyd med.

Til slutt kan vi si at både logistisk regresjon og k-nærmeste-nabo nabo fungerer godt, men at førstnevnte kan oftere gi bedre prediksjoner.

Oppgave 3: Klyngeanalyse

Q3.1:

a) *Hvor mange observasjoner (n) og hvor mange variabler (p) har vi?*

I kyngeanalyse har vi en datamatrise med n rader, observasjoner, og p kolonner, variabler. I dette tilfellet har vi n = 273280 observasjoner og p = 3 variabler. Her er n antall pixler og p er antall fargeverdier (rød, grønn og blå).

b) *Hvor finner du fargeverdiene til observasjonen med posisjon (x,y)=(10,20) i bildet i den nye tabellen "data_farger"?*

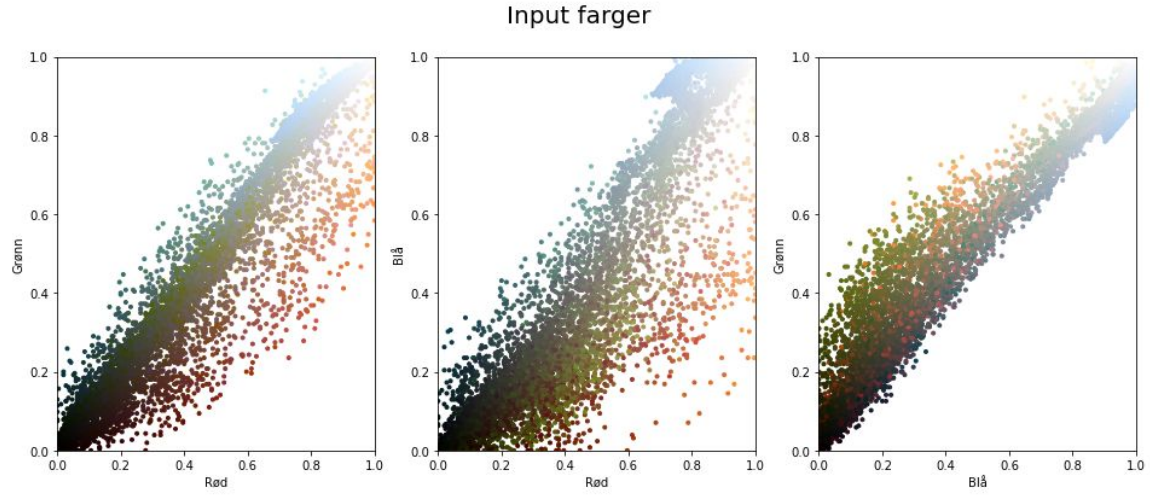
Hvis pixlene ligger radvis etter hverandre i den nye tabellen, må vi finne ut hvor mange elementer det er i hver rad og kolonne før den raden vi vil undersøke. Dette kan gjøres ved å gange antall elementer i hver kolonne med antall rader vi er ute etter fram til raden før den

kolonnen vi vil ha, da må vi gå til den 20. kolonnen i den raden. Vi husker at indeksene starter på 0:

$10 \times 640 + 1 \times 20 = 6420$ som er den indeksen vi er ute etter.

Q3.2:

Se kode under for hvordan lage et redusert datasett, og plotte rød mot grønn. Du legger til plott av rød mot blå og blå mot grønn. Kommenter hva du ser.



Et punkt = 1 pixel

Lave verdier av rød og grønn er sammen, noe man ser tydelig på bygningen i bildet. Vi ser at det er liten spredning av blå pixler, som kan forklares ved at blått ikke forekommer så ofte sammen med andre farger. Vi ser at himmelen og sjøen er blå, uten så mye innvirkninger fra andre farger.

Q3.3:

a) *Hva er sentroidene for ditt bilde?*

De 2 sentroidene med 2 klynger er

[0.82383625 0.85436035 0.88365767]

[0.28258494 0.2548509 0.18460248]

b) *I hvilken klynge havner fargene svart, hvit, rød, grønn, blå og gul?*

Svart, rød, grønn, blå: [0]

Hvit, gul: [1]

Tre viktige kodelinjer og så gjøres kryssplottene på nytt,

c) *Diskuter kort hva du ser.*

Vi ser at pixlene tilhører en av de to sentroidene vi fant over. Dette kommer tydelig frem ved at vi har to markante prikker i plottet som tilsvarer fargeverdiene til sentroidene vi fant.

Vi ser også at den første klyngen inneholder litt mindre blå enn rød og grønn, og at den andre klyngen er motsatt av det.

d) *Ved å se på det opprinnelige bildet, er det mulig å se hvilke deler av bildet som hører til hvilken klynge? Forklar!*

Vi kan finne ut av hvilke deler av bildet som hører til hvilken klynge ved å se på hvilken klynge fargene havner i, i oppgave 3.2 b) og se hvilke områder av bildet som domineres av hvilken farge. For eksempel vil deler med mye hvitt og lyseblått tilhøre klynge 1 og deler som domineres av mørke farger som svart, rød, grønn tilhøre klynge 0. Ved å se på det opprinnelige bildet ser vi at det er ganske todelt i skillet mellom forgrunn og bakgrunn. Man kan dermed se ganske tydelig ut i fra bildet hvilke deler som hører til hvilke klynge.

Q3.4:

Kommenter og forklar hva du observerer.

Vi ser først det opprinnelige bildet vårt, deretter ser vi bildet vårt etter 2-gjennomsnitt klyngeanalyse. Vi ser også at dette er fargene til sentroidene vi fant tidligere.

Neste vi ser er bilde vårt i svart hvitt. Her har klyngene fått henholdsvis svart og hvit farge.Til slutt ser vi to bilder med henholdsvis klynge 1 og 0 slått av.

Når vi slår av klynge 1 som består av lyse farger, ser vi at havet, himmelen og deler av taket blir fjernet fra bilde. Når klynge 0 blir slått av får vi motsatt effekt, hvor de mørke fargene i skogen og bygningen fjernes, mens havet og himmelen vises. Dette stemmer overens med diskusjonen vår om hvilke deler av bildet som tilhører hvilken klynge i oppgaven over.

Q3.5:

Hva er hovedforskjellene mellom K-gjennomsnitt-klyngeanalyse og hierarkisk klyngeanalyse? Vi ber deg ikke om å finne klynger i bildet ved hjelp av hierarkisk klyngeanalyse. Hva kan være grunnen til at vi ikke gjør det?

Hovedforskjellen på Hierarkisk- og K-gjennomsnitt klyngeanalyse er at i K-gjennomsnitt så definerer du antall klynger på forhånd. Man optimaliserer på klyngesentroidene, og finner en tilnærmet løsning ved å kjøre flere ganger med tilfeldig initialisering. I en Hierarkisk

klyngeanalyse så definerer man den maksimale avstanden for å koble sammen deler av klyngen. Vi ser at forskjellige avstander gjør at vi kan få forskjellige klynger. Dette er klynge hierarkiet vårt, som vi viser i et dendrogram.

Hierarkisk klyngeanalyse har høy kompleksitet, som gjør den uegnet for et stort datasett slik som vi jobber med. Det er derfor mer gunstig med K-gjennomsnitt.

Q3.6:

Hvor mange klynger trenger du for at du synes at bildet ser omtrent ut som det opprinnelige bildet? Prøv ut med ulike antall klynger og finn et klyngeantall du synes gir en god tilnærmelse, både med tanke på farger og detaljer. Hvor mange bit blir brukt per piksel i ditt valg av antall klynger over?

Ved få klynger ser man at bildet er veldig vannet ut og brunt i forhold til originalen. Vi valgte derfor å teste med høyere og høyere klynge tall. Som dere kan se av bildene under resulterte dette i bedre og bedre resultater. Allerede ved 11 klynger ser vi at fargene begynner å komme fram. Ved 16 klynger ligner bildet nok på originalen til at man kan se for seg hvordan bildet egentlig skal se ut. Vi valgte likevel å teste med høyere antall, og som man ser så gir både 22 og 55 klynger betraktelig bedre resultater. Dette kommer riktignok til prisen av lang ventetid når algoritmen kjører.

Når vi bruker 16 klynger betyr dette at vi får 16 forskjellige farger å jobbe med i bildet. Dette omtales ofte som “4-bit color”, og var ganske standard da VGA ble brukt. Det betyr at hver piksel bruker 4 bit.

4: 11: 15: 16: 22: 55:

Eirik Steira, Stian Mogen, Nicolay Schiøll-Johansen