2.Ökonometria házi feladat

ESS2020 adatbázis

Hajdu Bálint

VG8M5M

2023. 12. 04.

Az adatbázisról

A második házi feladatban az ESS2020 adatbázist kaptam, amely a European Social Survey magyar kitöltőinek adatait tartalmazza 1099 megfigyelésről. A megfigyelések 12 változó szerint vannak jellemezve. Ezek közül 1 a kitöltőkhöz rendelt ID, 3 mennyiségi és 8 minőségi változó. A kezdeti adatbázist a beimportálás és az átkonvertálás után ESS2020-nak neveztem el és ezzel dolgoztam tovább.

Hiányzó adatok kiszűrése

Az adatbázis megvizsgálása után találtam 26 darab hiányzó (NA) értéket. Ezek a hiá nyzó értékek 2 kategóriában merültek fel. Ebből 17 hiányzó értéket a PoliticalRadioTVPerDay _Minutes változóban, 9-et pedig az Education_Years változóban találtam. Az R-ben ezt kezelt em és amelyik sorban hiányzó értéket találtam, azt kiszűrtem a megfigyelésekből. Ennek ered ményeként 1073 megfigyelésem maradt.

Minőségi változók meghatározása

A hiányzó adatok kiszűrése után átalakítottam a bekezdésben említett 8 darab minős égi változómat. A TrustInParlament, SecretGroupInfluenceWorldPol, ScientistsDecievePublic , COVID19, ContactCOVID19, illetve a GetVaccince oszlopok adatait logikai változókká kon vertáltam, mivel Yes/No értéket vettek fel. A PoliticalPartyPref és Region változóimat pedig f aktorrá alakítottam, mert ezek kategória típusú adatokat vettek fel.

Outlierszűrés

Az előbb elvégzett lépések után léphettem a házi feladat második feladatpontjára az outlierek kezelésére. Ha az outlierszűrésnél nagyon szigorú akarok lenni és csak az adatok 1% -át engedem kidobni, akkor maximum 10-11 megfigyelés szűrhető ki az adatbázisomból. Ha e gy kicsit toleránsabb akarok lenni, akkor az adatok 2-3%-át is megengedhetem kidobni a megf elelő indoklással alátámasztva, ebben az esetben 32-33 adat kiszűrése is elfogadható lenne. En nél nagyobb százalékú és mennyiségű adat kistűrése csak a megfelelő nyomós indoklás esetén lenne engedélyezett és valóban megmagyarázható.

Az R-ben lefuttattam a summary függvényt az adatbázisra és elég érdekes adatokat k aptam mindhárom mennyiségi változónál. Ezeket egyesével vizsgáltam meg és mindegyiknél megnéztem a dobozábrákat. Először a PoliticalRadioTVPerDay_Minutes változómat vizsgálat ával kezdtem a vizsgálatot. Első gondolkodásra a dobozábra adatai relevánsnak tekinthetőek, mert benne vannak a 24 órás napi limitben (24*60=1440 perc). Az összes adat ezen korlát alat t van. Viszont, ha jobban megvizsgálom, akkor vannak elég érdekes eredmények is. Vannak o lyan emberek, akik 15+ órát hallgatnak politikai rádiót. Ez egy kicsit irrelevánsan tud hangzan i. Feltételezzük, hogy egy átlagos embernek 8-10 óra szükséges alvásra és emellett tegyük fel, hogy a maradék időt maximálisan kihasználja, amíg ébren van és egész nap a politikai rádiót h allgatja. Úgy gondolom, hogy ez releváns és hihető és ez azt jelenti, hogy 14-16 órát tud kihas ználni az ember. A dobozábrán látszik, hogy 800-nál van egy kis törés, ahol nincsenek adatok, ezért én ennél az értéknél húzom meg a racionális határt. A 800 perces határ nem adna vissza n ekem kerek egész értéket, ha ezt átkonvertálnám órába, ezért a kerekítés és a szép egész szám miatt 840 perces (14 órás) felső határt húzok meg. Így belefér az észszerűnek vélt és szüksége s 10 óra alvás is. Ez azt jelenti, hogy a 840 perc fölötti megfigyeléseket én kiszűröm ezen vált ozó mentén. Ez 27 adat kiszűrését jelenti és ezzel 1046-ra csökkent a megfigyelések száma. L efelé kilógó értékek nincsnek, viszont a 0 értékekből 111 van. Ezt két okból nem szűröm ki: ez rengeteg adat, de ami a legfontosabb, hogy ezek relevánsak és értelmezhetőek, mert rengetege n nem hallgatnak politikai rádiót.

A második vizsgált változóm a NetUsePerDay_Minutes volt. Ismételten az mondható el, mint az előző magyarázóváltozónál. Első gondolkodásra az értékek relevánsak, mert benne vannak a 24 órás/1440 perces napi korlátban. Viszont, ha jobban megvizsgálom, akkor itt is vannak érdekes, irracionális értékek. Tehát itt is nehéz elképzelni a 900 perc (15 óra) feletti értékeket. Itt is tegyük fel, hogy egy átlagos embernek 8-10 óra szükséges alvásra és a maradék időt maximálisan kihasználja és az internetet használja addig. Ez ismét 14-16 órát jelent, amit ki tud tölteni produktívan egy ember. A dobozábrán itt is látszik egy törés 800 percnél, ahol nincsenek adatok. Ezért ezen okok mentén itt ismételten a 14 órás (840 perces) határt húznám meg, szintúgy azon indokok mellett, mint az előző bekezdésben. Szóval a 840 perc feletti értékeket én kiszűröm. Ez 7 darab adatot jelent. Lefelé nincsenek kilógó értékek, viszont vannak alacsony értékek: pl.: 0,2,10. Ezek az értékek teljesen racionálisak, mert vannak olyan emberek, akik valóban keveset használják naponta az internetet. Egy lépést viszont most tennék meg, amely a házi feladatom későbbi szakaszában lesz releváns. (Kicsi spoiler.) A 0 értéket kiszűrném, mert az eredményváltozómat logaritmizálni fogom és a 0 érték nem tenné ezt nekem lehetővé úgy, hogy dolgozni tudjak a modellel. (Mivel a logaritmus feltétele, hogy ln a esetében a > 0). Ezt a problémát úgy is tudnám kezelni R-ben, hogy hozzáadok egy konstans számot a NetUsePerDay_Minutes értékeihez, viszont mivel csak 1 darab 0 érték van, ezt az egyszerűség kedvéért most inkább kiszűröm. Összegezve így 7+1, azaz 8 adatot fogok kiszűrni, amely 1038-ra csökkenti az adatbázisomban a megfigyelések számát.

A harmadik és egyben utolsó vizsgált mennyiségi változóm az Education_Years volt. A dobozábra ránézve itt is egy kicsit érdekes eredmények születtek. Van néhány alacsony érték, ami azt jelenti, hogy valaki még a 8 általánost évet sem végezte el. Ez a gondolatmenet egy darabig lehet racionális, viszont valahol én úgy gondolom meg kell húzni egy alsó határt. A 4 oktatási év azt jelenti, hogy valaki elkezdi 6 évesen az iskolát és 10 évesen abbahagyja. Ezt nehéz elképzelni. A dobozábrán mellesleg ez lefelé kilógó értékként látszik. Ezen indokok miatt ezt a 2 megfigyelést én itt leszűrném. Tovább gondolkodtam és vizsgáltam a 8 év alatti értékeket. A 6 és 7 éveket én racionálisabban fel tudom dolgozni. Például azt már könnyebben el tudom képzelni, hogy valaki 6 év oktatás után hagyja ott az iskolát 12 évesen ilyen vagy olyan egyéb indokok miatt. Ez is egy nagyon alacsony oktatási évszám, de ez már számomra elfogadhatóbb. Mellesleg a dobozábra alapján ezen értékek nem outlierek. Itt ezen indokok miatt az alsó határt a 6 évnél húzom meg. és csak az ennél kisebb adatokat szűröm ki. Amint már írtam, ez 2 adatot érint. Ezután megvizsgáltam a dobozábra felfelé kilógó értékeit. Felfelé is vannak érdekesebb adatok, de ezek számomra relevánsak. A 20-30 oktatási évet én azzal hihetőnek tudom magyarázni, hogy valaki elvégzi a 12 osztályt utána pedig akár 2 BSc diplomát és 1 MSc diplomát tesz le. Vagy akár orvosnak, jogásznak tanul és mellette más tanulmányokat is elvégez. Pl.: doktori cím megszerzése stb. Ezek időigényes folyamatok és ennek függvényében teljesen relevánsak nekem az adatok a 30 oktatási évig. A dobozábrán látottak alapján van egy jobban kilógó érték, a 40 év. Úgy gondolom ez már nagyon torzítaná a modellt és kicsit nehezebben is tudom ezt már racionálisan feldolgozni. Az még hihető és racionális volt számomra, hogy az említett indokok alapján valaki a 26-36 életévei között fejezi be a tanulmányait (Tegyük fel, hogy 6 évesen kezdte el az 1. osztályt és nem bukott és szüneteltetett.) Az viszont már számomra irracionális, hogy valaki 40 éven keresztül tanul és 46 éves korára fejezi be a tanulmányait (Itt is feltéve, hogy 6 éves korától oktatásban részesül.) Esetleg úgy is fel lehet dolgozni az adatokat, hogy valaki egész élete során tanul és a 80-90 életévéből szünetekkel együtt 20-30-40 évet töltött oktatásban. Úgy gondolom a 40 év itt is látható, hogy majdnem egy ember életének a felét jelentené. Szóval ezen gondolkodásmenetek mentén én itt

a 30 évnél húznám meg a felső határt és kiszűrném az ennél nagyobb adatokat. Ez 5 adat kiszűrését jelentené. Ezen változó mentén tehát 2+5, azaz 7 megfigyelést szűrnék ki az adatbázisból, amellyel véglegesen 1031-re módosul a megfigyelések száma.

Összegezve tehát az outliervizsgálatot a 3 mennyiségi változó mentén végeztem el. Ezek között 27+8+7=42 adatot szűrtem ki. Ez sajnos meghaladja mind az 1%, mind a 2-3%-os maximálisan leszűrhető megfigyelések számát, amelyet a bekezdésem elején határoztam meg. Véleményem szerint itt létfontosságú volt ezen adatok leszűrése, hogy ne kapjak torz becsléseket és eredményeket a házi feladatom során. Én úgy gondolom, hogy a gondolatmenetem és az indoklásaim észszerűek és relevánsak voltak és hiába léptem át a maximálisan tolerálható limitet az outlierszűrésnél, szerintem így is jogosan tettem ezt.A dobozábrákra utólág ránézve látható, hogy jól határoztam meg a határokat, amiken keresztül kiszűrtem az adatokat. Összesen tehát a kezdeti adatbázisom 1073 megfigyeléssel rendelkezett.(Ebből 26 megfigyelést hiányzó értékek miatt már kiszűrtem). 42 megfigyelést ebből outlierszűrés miatt távolítottam el és így 1031-re csökkentettem az adatbázisom, amellyel dolgozni fogok. Ami azt jelenti, hogy az adatbázisom 96,1%-át megtartottam és 3,9%-át hagytam el outlierek miatt.

Leíró statisztika a mennyiségi változókra

A házi feladatnak ugyan nem kötelező része, viszont elkészítettem egy leíró statisztikai elemzést a 3 mennyiségi változómra. Ennek eredményei itt láthatóak.

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
PoliticalRadioTVPerDay_Minutes	1031	53,12318	100,093	30	35,12606	29,652	0	792	792	5,372056	31,19457	3,117269
NetUsePerDay_Minutes	1031	175,7701	129,0431	120	154,0873	88,956	2	780	778	1,654975	2,965923	4,018884
Education_Years	1031	12,8322	2,920961	12	12,65697	1,4826	6	30	24	0,942767	1,741979	0,09097

PoliticalRadioTVPerDay_Minutes	NetUsePerDay_Minutes	Education_Years
Min. : 0.00	Min. : 2.0	Min. :6.00
1st Qu.: 15.00	1st Qu.: 90.0	1st Qu.:11.00
Median : 30.00	Median :120.0	Median :12.00
Mean : 53.12	Mean :175.8	Mean :12.83
3rd Qu.: 60.00	3rd Qu.:240.0	3rd Qu.:15.00
Max. :792.00	Max. :780.0	Max. :30.00

Az adatokat nem fogom egyenként, mivel nem a házi feladat része és ezért csak érdekességeket fogom kiemelni. Az ilyen fontos dolog, amire ki szeretnék térni, azok az adatokból látható eloszlások. Mindhárom mennyiségi változónál látszik, hogy az hozzájuk átlag magasabb, mint medián. Ebből következik, hogy mindhárom változó balra ferde, jobbra elnyúló eloszlású lesz. Ez a megállapítás látható abból is, hogy mindhárom változó ferdeségi mutatója pozitív. Ez már egy most fontos információ volt számomra, innen is már sejtettem, hogy az eredményváltozómat majd logaritmizálnom kell, hogy normális eloszlást kövessen. Nem meglepő ez a balra ferde eloszlás, mivel a változók alulról korlátosak és nagy mennyiségben kisebb értékeket vesznek fel az adatok. Ezek az információk tehát nem voltak annyira meglepőek nekem. Az eredményváltozó csúcsossági mutatójából látszik, hogy a normálishoz szinte hasonló a csúcsossága.

Korrelációs mátrix vizsgálata

Ez bekezdés sem a házi feladat része, viszont a teljesség képe miatt úgy gondolom fontos említést ejteni a korrelációs mátrixról. Láthatóak lesznek benne azok a várakozásaim, hogy a változók között valóban nagyon gyenge kapcsolatok vannak. Miután lefuttattam az R-ben a kódokat, azt kaptam, hogy minden kapcsolat nagyon gyenge és $|r| < 0,1$. Ez nem volt

meglepő számomra, egyedül arról nem volt elképzelésem, hogy ez az apró kapcsolat milyen irányú lesz. A NetUsePerDay_Minutes és a PoliticalRadioTVPerDay_Minutes, illetve a NetUsePerDay_Minutes és az Education_Years között nagyon gyenge pozitív/egyirányú a kapcsolat. A PoliticalRadioTVPerDay_Minutes és az Education_Years között viszont egy nagyon gyenge negatív/ellentétes kapcsolat áll fent.

Többváltozós lineáris regressziós modell és modellspecifikáció tesztelése

Először is, mielőtt lineáris regressziót futtattam volna, megnéztem a 2 faktor változóm referenciacsoportját. A PoliticalPartyPref változónál az Egyéb volt a viszonyítási alap. Ezt átállítottam Fidesz-KDNP-re, mivel ez a párt van jelenleg hatalmon és ezért nekem észszerű ehhez hasonlítani. Mellesleg ezt könnyeben lehet értelmezni, mint az Egyéb kategóriát. A másik változómnál, a Region-nél a Budapest/Pest a viszonyítási alap. Ez nekem megfelel, mivel általában az ország más pontjait ehhez szokták hasonlítani, mivel ez van középen és ez az ország központja.

A lineáris regressziós modell lefuttatása után megvizsgáltam az eredményeket. Ez egy elég gyenge magyarázóerejű modell. Az R^2 az 9,51%., azaz az eredményváltozóban lévő információnak csak a 9,51%-át magyarázza a modell. A korrigált R^2 az 8,54%. A Globális F-próba p-értéke nagyon alacsony, ez azt jelenti, hogy minden szokásos szignifikanciaszinten elutasítjuk a H0-t. Ez azt jelenti, hogy nem csak a mintában, hanem a való életben, a sokaságban is gyengén magyarázzák a változók a napi internethasználatot. Röviden, a modell kiterjeszthető a sokaságra. Amikor ránéztem az excel adatbázisra, akkor őszintén elcsodálkoztam az adatokon és azon, hogy tartalmilag mennyire eltérő változókat tartalmaz. Nem számítottam nagy magyarázóerőre, maximum 10-15%-os R^2-et vártam, azt sejtettem, hogy talán éppen megüti a közepes magyarázóerő határát (, ami 10%-tól kezdődik). Az eredmények majdnem elérték a határt, de végül nem tudták meglépni ezt. Elsőre nagyon alacsonynak tűnhet az eredmény, de egyáltalán nem meglepő. Összességében azt kaptam, amire számítottam, hogy ez a modell gyenge magyarázóerejű lesz és ez a sejtésem be is igazolódott. Úgy gondolom, hogy a napi internethasználatot jól lehetne magyarázni, olyan változókkal, mint például telefon képernyőideje, laptop vagy számítógép használata percekben vagy éppen TV nézése percekben. Amikor megláttam az eredményváltozót ilyen változókra gondoltam, amik szerintem jól tudnák magyarázni a napi internethasználatot. Mivel az ilyesmi és hasonló változók hiányoztak az adatbázisból, egyáltalán nem meglepő az eredmény ezért sem számomra.

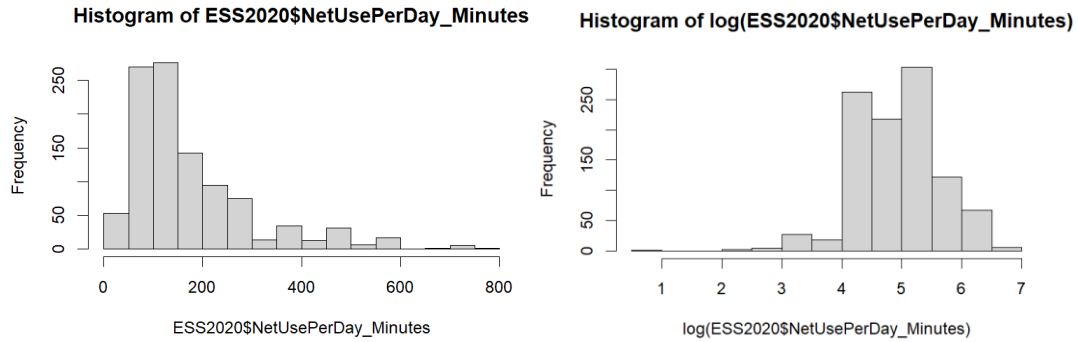
A modellnek a specifikációját is teszteltem Ramsey-féle RESET-tesztel. Amelynél a H0: a modell jól specifikált, H1: a modell nem jól specifikált - null és alternatív hipotéziseket vizsgáltam. A teszt eredménye 5,38%-os p-értéket adott vissza, szóval H0-t nem utasítjuk el minden szokásos szignifikanciaszinten. 10%-on még elutasítjuk, de Alfa= 5% és 1%-os szignifikanciaszinteken még elfogadjuk. Ez azt jelenti, hogy nem minden szokásos szignifikanciaszinten mondható el, hogy a modell jól specifikált. Ahhoz, hogy minden szokásos szignifikanciaszinten jól specifikáltnak mondható legyen, bővíteni kellene a modellt a változók interakcióival vagy/és nem-lineáris transzformáltjaival.

Mielőtt az elkezdtem keresni az interakciókat és a nem-lineáris specifikációkat, azelőtt kidobtam az összes nem szignifikáns változómat a PoliticalRadioTVPerDay_Minutes kivételével. Ezt a változómat azért nem dobtam ki, mert ha ezt kivenném, akkor 1 darab mennyiségi változóval maradnék a főkomponens elemzéshez és ezt el akartam kerülni. A modell_1_szukitett-ben tehát létrehoztam az említett adatbázist és az információs kritériumok

és a Wald-teszt alapján is azokat az eredményeket kaptam, hogy a szűkített modell a preferált. Ezek után már nyugodtan kezdhettem neki az új tagok keresésének.

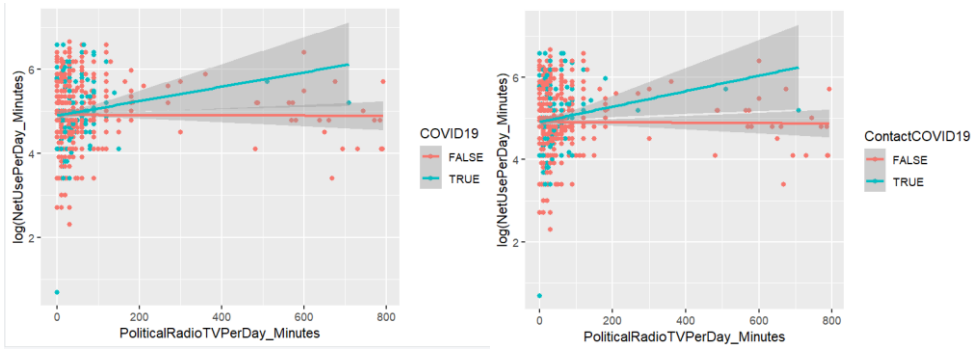
Interakciók, nem-lineáris specifikációk keresése

A házi feladatban a következő lépés a modell bővítéséhez szükséges interakciók és nem-lineáris specifikációk vizsgálata volt. Ezen vizsgálatot én hosszasan elkészítettem az R-ben és én úgy gondolom, hogy a legrelevánsabb kombinációkat én megvizsgáltam. Itt a dokumentumban én csak azokról teszek említést, amelyek relevánsak vagy éppen érdekességek. Először a mennyiségi változók eloszlásait vizsgáltam meg hisztogramon. Az eredményeket, viszont már előre tudhatom, hogy mindhárom változóm balra ferde, jobbra elnyúló lesz a leíró statisztika miatt. A modellfeltevés feltétele ugyan teljes a nagy minta esetével, de jobb eredményeket kapok akkor, ha az eredményváltozóm normális eloszlást követ. Ezen indok mentén a hisztogrammal alátámasztva én logaritmizálni fogom az eredményváltozómat. A hisztogram eredményei itt láthatóak az eredményváltozómra.



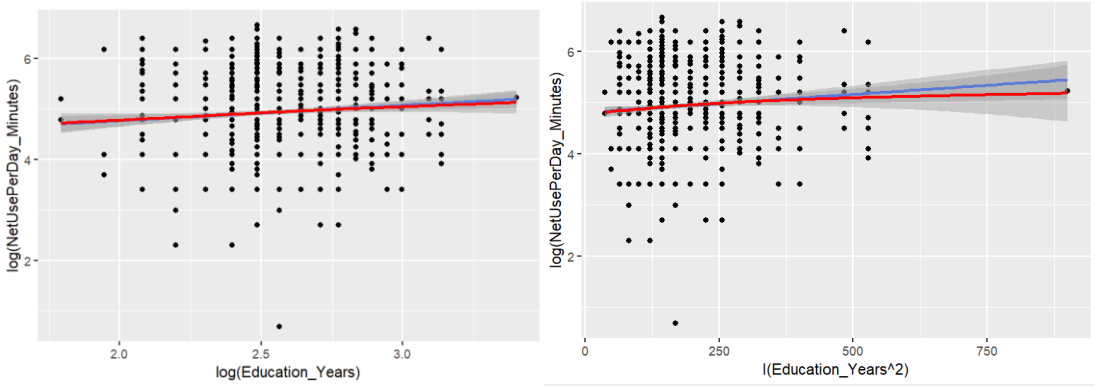
Úgy gondolom a logaritmizált értékek pontosabban követik a normális eloszlást és amint már említettem, az eredményváltozó értékei alulról korlátosak és nagy mennyiségben az adatok a kisebb értékek körül tömörülnek, ezért érdemes logaritmizálni. A másik két mennyiségi változómra szimplán a hisztogram alapján nem tudom eldönteni, hogy a logaritmizálás lenne-e a legmegfelelőbb opció.

Következőnek a lehetséges interakciókat vizsgáltam meg mindkét mennyiségi magyarázóváltozóm esetében. Itt 2 eset kivételével nem találtam drasztikus különbséget az egyenesek meredekségei között és így arra jutottam, ahol nem volt akkora különbség a meredekségek között, hogy ott nem kell interakciót alkalmazni. A 2 eset, ahol én lehetségesnek tartom az interakció bevételét a modellbe, az a PoliticalRadioTVPerDay_Minutes és a COVID19, illetve a PoliticalRadioTVPerDay_Minutes és a ContactCOVID19 interakciója. Ezek a meredekségkülönbségek így néznek ki, viszont ez a kilógó értékek miatt is alakulhatott így, ezért nem tudom biztosan eldönteni, hogy valóban jó lenne-e az interakciók bevételére.



Ezeknek az interakcióknak a bevételéről azután döntök, hogy megvizsgáltam minden kapcsolatot.

Az interakcióknak a bevétele után én a mennyiségi magyarázóváltozóimat vizsgáltam tovább, hogy érdemes-e belevenni a logaritmizált vagy éppen a négyzetes tagjukat. Az Education_Years esetében az ábra alapján mind a logaritmizált, mind a négyzetes értéke bevehető lenne a modellbe. Azoknál az adatoknál, ahol évek vannak nagyon gyakran a négyzetes tag bevétele adja a legjobb eredményt számunkra, szóval már emiatt inkább ennek a bevételére hajlottam, de azt, hogy pontosan a kettő közül melyik a jobb, azt az interakciókkal együtt közösen döntöm majd el. Itt láthatóak a logaritmizált és a kvadratikus kapcsolatok. Ránézésre nehéz eldönteni, hogy melyik eredményezne jobb eredményt.



Ezután megnéztem a PoliticalRadioTVPerDay_Minutes változóm kapcsolata hogyan néz ki az eredményváltozóval, ha logaritmizálva van. Itt, ha logaritmizálom a magyarázóváltozómat, akkor muszáj előtte egy konstans összeget hozzáadni, mivel rendelkezik 111 darab 0-ás értékkel. Az kapcsolatot megvizsgáltam különböző konstans hozzáadott értékekkel és arra jutottam, hogy a konstans értékétől függ az, hogy hogyan néz ki a pontdiagramm. Az pedig, azt jelentené, hogy el kellene döntenem, hogy mi az a konstans érték, amivel növelem a változómat logaritmizálás előtt. Ennek okán én itt úgy döntöttem, hogy nem logaritmizálom a változómat és a modellemben benne hagyom a sima változót bármiféle transzformáció nélkül.

Az utolsó bekezdésben pedig azt szeretném bemutatni, hogy végül milyen új tagok bevétele mellett döntöttem. Az R-ben 7 darab próba modellt futtattam, annak érdekében, hogy megvizsgáljam azt, hogy az interakciókat érdemes-e valóban bevenni, az Education_Years a logaritmizált értékek vagy a négyzetes értékek lennének-e jobbak. Emellett azt is szemléltettem, hogy a PoliticalRadioTVPerDay_Minutes esetében a konstans megváltoztatása a logaritmizálásnál valóban befolyásolja az eredményeket. A modellek lefuttatása után

megkaptam tehát azt, hogy az Education_Years esetében jobb, ha a négyzetes értékeket veszem, mert ebben az esetben jobb lesz a változóm szignifikanciája és a modell specifikáltsága. Az interakciók esetében, pedig azt kaptam, hogy káros hatással vannak vagy a PoliticalRadioTVPerDay_Minutes változó szignifikanciájára vagy a modell specifikáltságára. Ezen okok mentén én tehát amellett döntök, hogy az interakciókat nem veszem bele a modellembe.

Összegzésként tehát elmondhatom, hogy a az alapmodellemet, a modell_1-et az eredményváltozóm logaritmizálásával és az Education_Years kvadratikus tagjával bővítettem. Ez lett a végső modellem, amire a következő bekezdésben újra elvégzem a RESET-tesztet.

Végső modell modellspecifikáció tesztelése

Amint az előző bekezdésben olvasható volt, sikerült megkapnom a végső modellemet, amelyben csak a szignifikáns változóimat tartottam meg, a PoliticalRadioTVPerDay_Minutes kivételével emellett az eredményváltozómat logaritmizáltam, az Education_Years változónak pedig a négyzetes értékeit vettem bele. Ezt a modellt az R-ben modell_2-nek neveztem el. A Ramsey-féle RESET-teszt elvégzése után azt az eredményt kaptam, hogy a p-értékem 25,57%-os, azaz jól specifikált. Ezzel javítottam az eredeti modellemen, amely azt mutatja, hogy jól dolgoztam.

Végső modell egyik marginális hatás értelmezése

Mivel az eredményváltozóm logaritmizálva lett, ezért ezt figyelembe kell vennem a béták értelmezésénél és emellett arra is figyelnem kell, hogy melyik változót értelmezem. A PoliticalRadioTVPerDay_Minutes változóm lineáris maradt, szóval ennek az értelmezése során én egy Log-Lin értelmezést kell figyelembe vegyek. A változóm béta értéke 0.0002312 lett. Ezt e-ad hatványra kell emelnem, hogy megkapjam a valódi változást. Ez azt jelenti, ha a Politikai rádióhallgatás egy nap 1 perccel nő, akkor a napi internethasználat a 1.000231-szeresére változik. Nem egy nagy változás, de ez volt várható egy ehhez hasonló változótól. Nem meglepő tehát az eredmény.

Főkomponens elemzés

Miután az R-ben kiszámoltam a VIF-mutatókat, megvizsgáltam azokat és azt vettem észre, hogy nincsen multikollinearitás a modellemben. Minden változóm VIF-mutatója bőven az 5-ös érték alatt volt, sőt az összes 1,1-es alatti értéket vett fel. A zavaró multikollinearitásról VIF-mutató>5, a káros multikollinearitásról pedig VIF-mutató>10-nél lehetne beszélni. Ez egyértelműen látszik, hogy egyik változó esetében sem merül fel, szóval szó sincsen multikollinearitásról. Ezen indokok mentén nincsen értelme főkomponenseket vizsgálni, mivel nem kell a modellt megtisztítani a multikollinearitástól.

A házi feladat teljessége érdekében viszont elvégeztem az R-ben a főkomponens elemzést és azt kaptam, hogy sajátérték alapján csak a PC1-et kellene bevennem a modellbe, de a kumulált információmegőrzés alapján, pedig mindkettőt bele kellene vennem, mert a PC1 ezen téren csak 51% információt őriz meg. A főkomponenseket tehát nem kell alkalmazni multikollinearitásra, mivel ez a jelenség nem áll fent.

Heteroszkedaszticitás vizsgálata

A házi feladatnak ezen bekezdés sem része, viszont a teljesség képe miatt ezt a részt is elkészítem. A hibatagok négyzetének az ábrázolása után látható, hogy az értékek nem véletlenszerűek. Ennek következtében itt fennáll a heteroszkedaszticitás jelensége. Ezután a Kolmogorov-Smirnov teszttel a hibatag eloszlásának a megvizsgálása után észrevettem, hogy nem normális eloszlást követ, tehát a Breusch-Pagan Koenker korrekciós verziója fogja visszaadni a legpontosabb eredményt a heteroszkedaszticitást tesztelését illetően. A teszt elvégzése után azt tudom mondani, hogy alfa=1%-os szignifikanciaszinten fennáll a heteroszkedaszticitás. A White teszt, mivel normális eloszlást feltételez, ezért azt adta vissza eredményül, hogy nincsen heteroszkedaszticitás. Ugyanaezt az eredményt kaptam a Breusch-Pagan teszt Koenker korrekciós verzió nélküli esetben is.

Összegzés

Az előfeltevéseim igaznak bizonyultak, miszerint a modellem nem fog erős magyarázóerővel rendelkezni, mivel nem rendelkezik jó magyarázóváltozókkal. Ezt igazolták az alacsony R^2 értékek is. A modellt megfelelően szűkítettem le, ezt igazolják az infromációs kritériumok és a Wald-teszt is. Majd ezek után a megfelelő nem-lineáris specifikációkat vettem bele a modellbe, ezt igazolja az is, hogy a végső modellem jól specifikált lett. Ezután megvizsgáltam, hogy a modellem nem tartalmaz multikollinearitást és ezáltal nincsen szükség főkomponensekre és főkomponensek bevételére sem. Legvégső esetben pedig megvizsgáltam a heteroszkedaszticitás jelenségét, és a megfelelő teszt megválasztása után arra az eredményre jutottam, hogy fenáll ez a jelenség.