KU LEUVEN kulak

Informatica
Statistische modellen en data-analyse
Prof. Van Aelst, Stijn Rebry

Opdracht 2

Academiejaar 2018 – 2019

Thomas Bamelis, Michiel Jonckheere

Inleiding

In dit verslag analyseren we de gegevens van airbnb gemeten in 2019 van de steden Antwerpen, Brussel en Gent. De data bevat in totaal ongeveer 11.000 samples. We analyseren de prijs in functie van het type, de stad en de buurt van het bedrijf. Daarna proberen we een regressiemodel op te stellen voor de huurprijs en dit model te evalueren en te interpreteren. Als laatste passen we logistische regressie toe om te voorspellen of een verblijf al dan niet wordt vast verhuurd in plaats van enkel voor korte periodes.
Opmerking: als in dit verslag een p-waarde als nagenoeg 0 wordt beschreven, betekent dit dat de p-waarde voor underflow zorgde in het computersysteem ($<$ 2.2e-16). Indien het significantieniveau niet vermeld staat wordt 0.05 gehanteerd.
De volgende gegevens zijn beschikbaar:

1. id
2. naam
3. uitbater id
4. naam uitbater
5. buurt
6. latitude
7. longitude
8. type verblijf
9. huurprijs
10. minimum aantal nachten
11. aantal reviews
12. datum laatste review
13. aantal reviews per maand
14. aantal verblijven/zoekertjes van de uitbater
15. hoeveel dagen het verblijf beschikbaar was afgelopen jaar
16. stad
17. of een verblijf al dan niet wordt vast verhuurd in plaats van enkel voor korte periodes

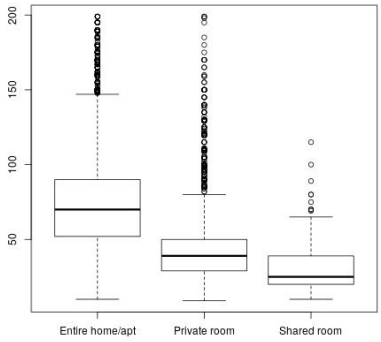
Deze laatste werd zelf afgeleid uit het aantal beschikbare dagen.

1 Ligging en type verblijf

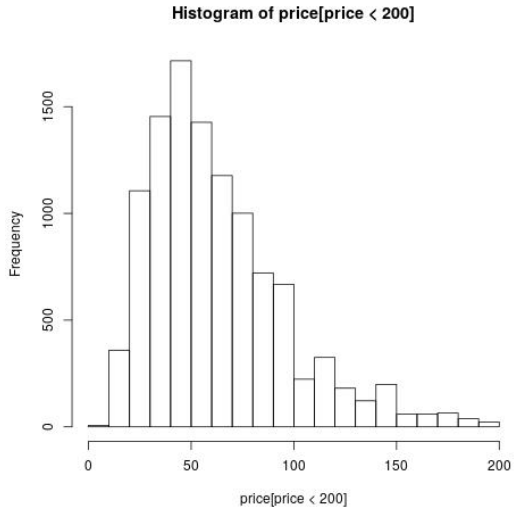
In deze sectie analyseren we de prijs in functie van het type, de stad en de buurt van het bedrijf. We analyseren of de stad nog significant is als de buurt wordt inbegrepen en de buurt als het type verblijd wordt meegerekend. Deze laatste omdat in bepaalde buurten misschien hoofdzakelijk een bepaald soort type verblijf aanwezig is.

1.1 type verblijf vs huurprijs

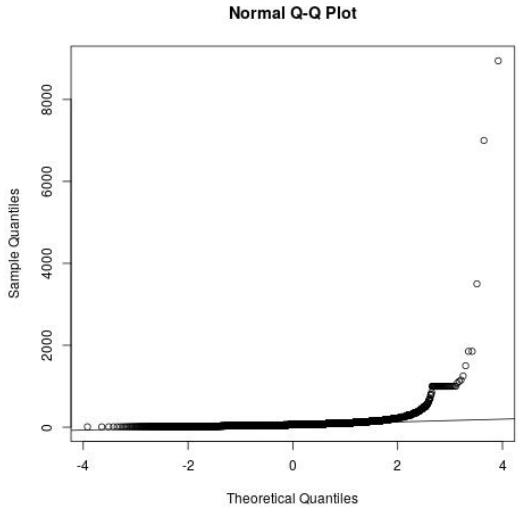
Een boxplot van de prijs tegenover de stad (1) zonder de outliers lijkt te suggereren dat er wel degelijk een verschil blijkt te zijn tussen de types van de verblijven, en dan vooral tussen een volledig huis/appartement en de rest. Om overzicht te behouden hebben we de bovenste outliers uit de figuur gelaten, waarvan er nog 705 tussen 150 en 1500 lagen en nog 8 hoger dan 1500 met een maximum van 8944. We analyseren eerst de verdeling van de prijs om beter over het komend werk te kunnen redeneren. De prijs lijkt aan de hand van figuur 2a een redelijke klokcurve voor te stellen met echter een zeer lichte linkerstaart en een heel erg zware rechterstaart. Opnieuw werden de grote outliers niet in de figuur opgenomen. Een normale kwantielplot van de prijs lijkt totaal niet normaal verdeeld door de enorm zware outliers en de knik in de curve. Een schatting van de lambda voor een Box-Cox transformatie raad -0.2549 aan als lambda, waardoor we een transformatie van -1/4 toepassen op de prijs. Het resultaat is te zien in figuur 2d, wat een zeer duidelijke verbetering toont op het vlak van normaliteit. Deze transformatie gaf de beste verbetering van alle geteste transformaties (log, sqrt, loglog).



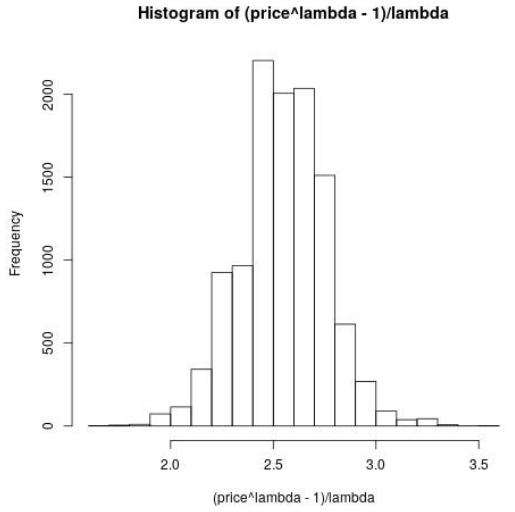
Figuur 1: Boxplot type vs prijs.



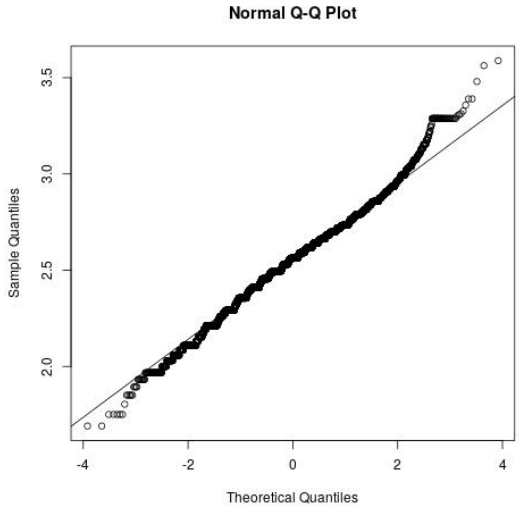
(a) Visualisatie van de prijs.



(b) Kwantielplot prijs



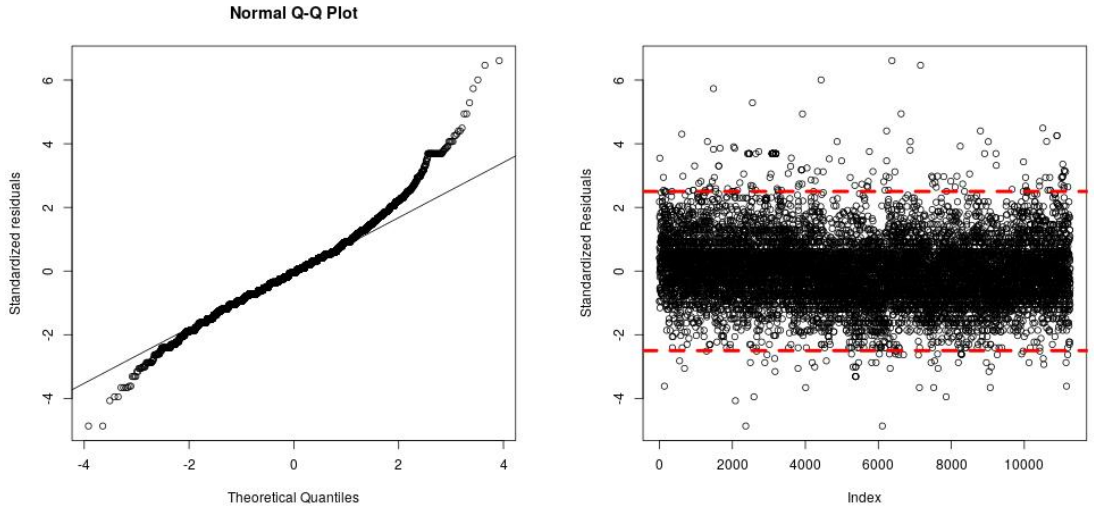
(c) Visualisatie Box-Cox -1/4 prijs



(d) Kwantielplot Box-Cox -1/4 prijs

Na de residuals bekeken te hebben van de modellen waarbij transformaties van de prijs voorspeld worden aan de hand van het type, vonden we dat de Box-Cox transformatie er als beste transformatie uitkwam (figuur 3a). De Levene test toont aan dat er heteroscedasticiteit is van de varianties tussen de verschillende groepen met een p-waarde van nagenoeg 0, wat aannemelijk lijkt gegeven figuur 1. Een kwantielplot van de gestandardiseerde residuals toont ook te zware staarten (3a), waardoor we concluderen dat ze niet normaal verdeeld zijn. Het aantal samples is te groot om een Shapiro test op uit te voeren. Een plot van de residuals tegenover de index (figuur 3b) toont geen afhankelijkheid "in de tijd". Door de heteroscedasticiteit en het niet normaal zijn van de residuals zijn de modelveronderstellingen voor de relevante testen niet voldaan, onze bevindingen moeten daarom met een korrel zout genomen worden.
Het model verwierp de f-test (p-waarde nagenoeg 0) en t-test voor alle variabelen (p-waarde nagenoeg 0 voor alle categorieën). De R-squared en adjusted R-squared waren echter maar 0.268. De weighted least square methode toonde geen verbeteringen op het model na 1 iteratie. De anova methode toont alsook aan dat het type verblijf effectief significant is voor de prijs met een p-waarde van nagenoeg nul. De tukey-test toont verder aan dat de verschillende soorten kamers onderling ook genoeg verschillen van elkaar (alle combinaties nagenoeg 0).
Conclusie: het type verblijf is significant voor de prijs en de types verschillen onderling allemaal in prijs. Volgens de coëfficiënten van het model is de prijs als volgt gerangschikt: volledig huis app. > privé kamer

> gedeelde kamer. Deze bevindingen stroken met de vermoedens van de realiteit.

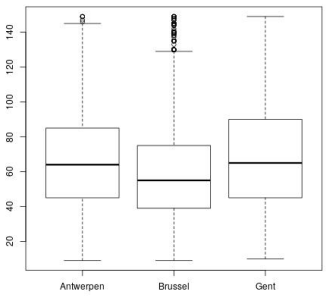


(a) Kwantielplot gestandardiseerde residuals

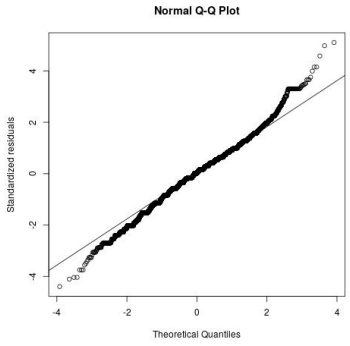
(b) Index vs gest. residuals

1.2 stad vs huurprijs

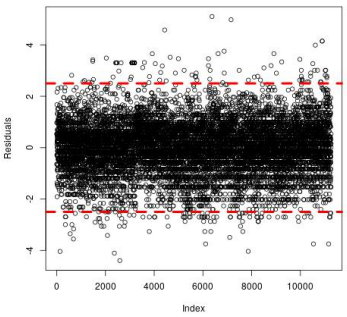
De boxplot zonder outliers (figuur 4a) suggereert dat er niet veel verschil is in de gemiddelde prijs per stad. Antwerpen en Gent zijn nagenoeg het zelfde, enkel de prijzen in Brussel liggen wat lager.
We nemen opnieuw dezelfde transformatie op de prijs zoals in sectie 1.1. We zien dat de residuals niet normaal verdeeld zijn (figuur 4b), waardoor onze bevindingen opnieuw niet steenhard mogen genomen worden. De Levene test kan echter niet verwerpen dat er homoscedasticiteit is op significantieniveau 0.05 met een p-waarde van 0.1454, waardoor de testen toch al serieuzer mogen worden genomen dan in sectie 1.1
Het model verwerpt de f-test (p-waarde nagenoeg 0) en t-test voor de variabelen van Antwerpen (= het intercept) en Brussel (p-waarde nagenoeg 0), maar Gent is niet verworpen op significantieniveau 0.05 met een p-waarde van 0.068. Er moet rekening meegehouden worden dat Antwerpen "voorang" krijgt aangezien het het intercept is, en de gelijkenissen tussen Antwerpen en Gent uit figuur 4a dit kunnen verklaren. De reden dat Gent dus niet verworpen is, is omdat hij heel sterk op Antwerpen lijkt, wat dus betekent dat beiden niet significant verschillen qua prijs. En het dus niet perse Gent die slechter is dan de andere twee. Het kleine verschil in de coëfficiënten bevestigd dat. Aov zegt dat de stad significant is voor de prijs op significantieniveau 0.05 met een p-waarde van nagenoeg 0. De residuals zijn niet afhankelijk van de tijd/index, dit volgt uit figuur 4c. De R-squared van dit model is 0.0226 en de adjusted is 0.0224, wat dus opnieuw niet goed is. De Tukey-test verwerpt sterk dat er geen verschil zou zijn tussen Brussel en de andere twee steden. Maar Gent en Antwerpen worden niet verworpen op significantieniveau 0.05 met een p-waarde van 0.162, wat opnieuw onze vermoedens bevestigd. Het effect van Brussel is ook veel sterker dan die van Antwerpen en Gent. Conclusie: de stad is significant voor de prijs. Er is een prijsverschil tussen Brussel en de andere 2 steden, maar er is geen significant prijsverschil tussen Antwerpen en Gent. De coëfficiënten van het model vertellen dat Brussel goedkoper is dan de andere 2 steden, wat strookt met de boxplot op figuur 4a.



(a) Boxplot stad vs prijs.



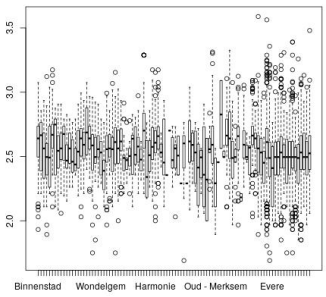
(b) QQplot gestandaardiseerde residuals model



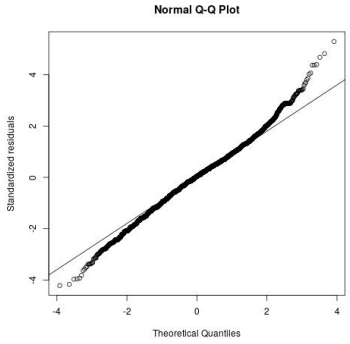
(c) Index vs residuals

1.3 Buurten

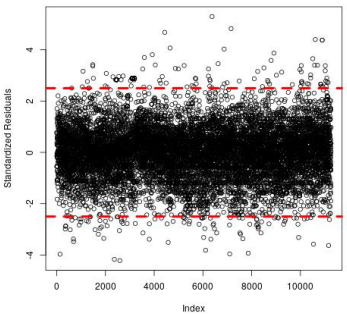
We nemen opnieuw dezelfde transformatie op de prijs zoals in sectie 1.1. Op het eerste zicht (figuur 5a) lijken de buurten niet zo veel te verschillen. Het model waarbij we enkel neighbourhood meenemen, verwerpt de f-test met p-waarde nagenoeg 0, maar een groot deel van de neighbourhoods worden niet verworpen door de t-test op significantieniveau 0.05 (43/96). Door de heteroscedasticiteit doe blijkt uit een Levene test en het niet normaal zijn van de residuals (figuur 5b) zijn de modelveronderstellingen voor de relevante testen niet voldaan, onze bevindingen moeten daarom met een korrel zout genomen worden. Residual zijn niet afhankelijk van de tijd/index (figuur 5c) Aov toont aan dat de buurt significant is met p-waarde nagenoeg 0. De TukeyTest kan 47,6% niet verwerpen, dus 47,6% van de combinaties van buurten verschillen niet significant. De R-squared is 0.099 en adjusted 0.091, wat opnieuw aan de lage kant is. Alleen lijkt dus dat de buurt significant is in het algemeen, maar zeer veel buurten onderling te weinig verschillen. Dit doet suggereren dat de opdeling per buurt te verfijnt is.



(a) Boxplot buurt vs prijs.



(b) QQplot gestandaardiseerde residuals model

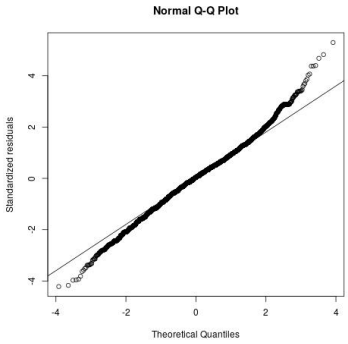


(c) Index vs residuals

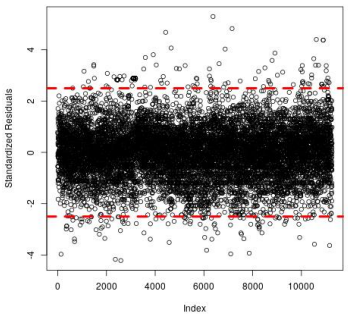
1.4 Buurten en stad

We nemen opnieuw dezelfde transformatie op de prijs zoals in sectie 1.1. Het model met de buurt en de stad verwerpt de f-test ook met p-waarde nagenoeg 0, maar een nog groter deel van wordt niet verworpen door de t-test (54). Meer bepaald is de pwaarde van Gent zeer hoog (0.9) terwijl Antwerpen (nagenoeg 0) en Brussel (0.016, wel hoger dus) nog steeds verworpen worden (Antwerpen omdat die het intercept is). De R-squared is 0.099 en adjusted 0.091 en blijven dus onveranderd bij het toevoegen van de stad. Dit doet vermoeden dat beide dus sterk verbonden zijn aangezien hun gecombineerd model even slecht is als hun afzonderlijke. Aov verwerpt echter voor beide variabelen en zegt dus dat ze beiden nog significant zijn. De Tukey test kan nu nog 44,6% van de combinaties verwerpen. Deze bevindingen suggereren dat door de buurt en de stad beiden op te nemen, de significantie van Gent vs. Antwerpen miniem wordt. De Aov verwerpt niet dat een van de twee overbodig wordt, maar aangezien zeer veel variabelen de t-test

falen, er geen verbetering is in de R-squared en het nog steeds hoge percentage van het niet verwerpen van de Tukey test tonen dat het origineel probleem van het te verfijnd zijn van de buurtenverdeling enkel nog maar versterkt wordt door de stad erbij te nemen.

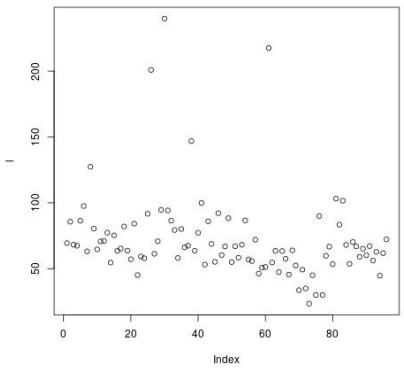


(a) QQplot gestandaardiseerde residuals model



(b) Index vs residuals

Figuur 7 toont dat er een ongeveer 5 buurten zijn die een opvallend hogere prijs hebben dan de rest. De gemiddelde huurprijs is 72,2 euro. Woluwe-Saint-Pierre steekt hier sterk boven met een gemiddelde huurprijs van 240 euro. Daarna volgt Sint Denijs Westrem met 217,5 en Oud-Berchem met 201. Deze drie zijn significant hoger dan de rest in liggen in Brussel, Gent en Antwerpen. Niet in een bepaalde stad dus. Dan zijn er nog 2 minder extreem prijzige gemeenten, namelijk Eilandje met 147 en Polder met 127 (beiden liggen in Antwerpen). De eerste opeenvolgende is nog maar 103. Daarnaast zijn er nog 5 Antwerpse buurten waar de prijs ietwat lager ligt dan de rest met prijzen tussen 23,5 en 35 euro. Deze zijn echter minder een stuk minder ver verwijdert van het gemiddelde in vergelijking met de extreem prijzige outliers.



Figuur 7: Gemiddelde prijs per buurt.

1.5 Buurten en type verblijf

Op analoge manier zijn we te werk gegaan om een model op te stellen van de buurten en het type. Opnieuw werd de f-test verworpen en 42 t-testen gefaald. Aov verwerpt echter voor beide variabelen en acht ze dus beide significant. Wat echter wel opmerkelijk is, is dat de R-squared sterk gestegen is van 0.099 naar 0.3281 en de adjusted van 0.091 naar 0.3222. De R-squared van room-type alleen was ook 0.2678 en de adjusted 0.2677. De combinatie van de 2 variabelen versterk elkaar dus. De buurt en het type verblijf maken elkaar dus niet overbodig.

2 Model voor de huurprijs

Op basis van relevante beschikbare gegevens hebben we geprobeerd een model op te stellen om de prijs van een Airbnb-verblijf zo goed mogelijk te voorspellen. Eerst moesten we daarvoor kijken welke gegevens er relevant zouden zijn voor ons model. Daarna controleerden we of het nodig was om bij bepaalde variabelen een transformatie toe te passen. We bekeken enkele modellen met en zonder transformaties en tot slot bepaalden we het beste model om de prijs te voorspellen.

2.1 Selectie relevante gegevens

Om variabelen te selecteren als een relevante variabele voor ons model, keken we naar hoe de prijs zich verhoudt ten opzichte van de variabele. Eerst en vooral merken we op dat ”id”, ”name”, ”host_id” en ”host_name” niets te maken kunnen hebben met de prijs, deze gegevens laten we dus achterwege. De variabele ”latitude” blijkt wel een in vloed te hebben op de prijs. Als we kijken naar de waarden van de breedtegraden dan zien we dat we deze kunnen indelen in de drie steden. We gaan in plaats van de breedtegraad in ons model te gebruiken, gebruik maken van de variabele ”city”. ”Longitude” wordt om dezelfde reden niet gebruikt in het model, deze variabele deelt de data op in twee delen namelijk als eerste deel Gent en als tweede deel Brussel en Antwerpen. Vervolgens zien we ook dat variabelen ”room_type”, ”minimum_nights”, ”number_of_reviews”, ”last_review”, ”reviews_per_month”, ”availability_365” en ”calculated_host_listings_count” relevant kunnen zijn voor ons model.

Voor de variabelen die iets zeggen over de reviews hebben ook eens gekeken naar de correlatie tussen deze variabelen. We vonden dat er een sterke correlatie bestaat tussen ”reviews_per_month” en ”number_of_reviews”, ook is er een correlatie tussen ”reviews_per_month” en ”last_review”. Om deze reden gaan we zeker al twee verschillende modellen opstellen waarbij enerzijds gewerkt wordt met alle drie de variabelen en anderzijds enkel met de variabele ”reviews_per_month”.

2.2 Transformaties

In deze sectie bekijken we of er transformaties zijn die ervoor kunnen zorgen dat we een beter model verkrijgen om de prijs te voorspellen.

Variabele	**Transformatie**
room_type	/
city	/
price	$\frac{(price^{(-0.25)})-1}{-0.25}$
minimum_nights	$\frac{(minimum\_nights^{(-0.67)})-1}{-0.67}$
number_of_reviews	log10(number_of_reviews)
last_review	log10(last_review + 1)
reviews_per_month	log10(reviews_per_month)
calculated_host_listings_count	$\frac{(calculated\_host\_listings\_count^{(-1)})-1}{-1}$
availability_365	/

2.3 Modellen

commentaar thomas:
variabelen selectie: teveel variabelen laten vallen in het begin, nu deze variabelen gebruiken: roomtype, price, minimumnights, nbreviews, lastreviez, reviewpermonth, calchost, city, availability
transformaties
interactietermen

model evalueren:
checken op multicolinearitеit checken
waar liggen de outliers
inferenties
5plots (pg216, residualplots)

3 Beschikbaarheid van een verblijf

In deze sectie proberen we logistische regressie toe te passen om te voorspellen of een verblijf al dan niet wordt vast verhuurd in plaats van enkel voor korte periodes. We proberen een model te vinden, dit te evalueren en te interpreteren.

3.1 Model selectie

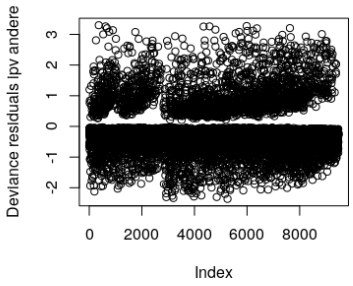
Voor we beginnen met een model te zoeken verwijderen we alle rijen die een kolom missen. Het aantal samples gaat dan van 11262 naar 9476. De proporties van categorische veranderlijken worden hierdoor niet significant aangepast. Het aantal beschikbare dagen wordt natuurlijk niet opgenomen als regressor omdat het al dan niet permanent verhuurd zijn direct hieruit is afgeleid. We laten de naam en id van het zoekertje en de host eveneens achterwege omdat id willekeurig zijn voor de andere attributen van een zoekertje. De naam (of id) van het verblijf is zinloos omdat dit een 1 op 1 relatie is per zoekertje en dit dus gigantische overfitting zou zijn. De naam van de host (of hun categorische id) zou in principe kunnen meegenomen worden, maar dit doen we niet om 3 redenen:

- kan gezien worden als sterke overfitting (zie puntje 3)
- het model wordt op die manier onoverzichtelijk
- het model heeft op die manier geen enkele betekenis meer als een nieuw verhuurder op de site komt en heeft dus geen voorspellende kracht om te voorspellen.

We pasten AIC gebaseerde voorwaartse, achterwaartse en stapsgewijze regressie toe. Dit deden we 2 maal, de 2e keer met de variabelen in de omgekeerde volgorde. Hierbij werden de prijs, het type, het minimum aantal nachten, het aantal zoekertjes per host, de datum van de laatste review, het aantal reviews per maand en het aantal reviews van het afgelopen jaar telkens geselecteerd. De buurt werd nooit geselecteerd, de stad 2 keer en de latitude en longitude 4 keer. Stad werd enkel geselecteerd als latitude en longitude niet geselecteerd werden en omgekeerd, wat dus suggereert dat ze dezelfde informatie toevoegen aan het model, wat ook strookt met de realiteit van de soort variabelen. Stad is de nette classificatie van latitude en longitude, daarom vergelijken we een model met latitude en longitude tegenover een model met de stad in de plaats. De buurt laten we achterwege. De bekomen modellen hebben dezelfde AIC en deviance waarden en resulteren in hetzelfde voor de likelihood ratio test en de goodness of fit test, behalve dat latitude maar zwak verworpen wordt in gun eigen model. Omdat de stad de significantie van latitude en longitude mooi samenvat gaan we verder met de stad en laten we latitude en longitude vallen. In de testen wordt echter niet verworpen dat het type er niet toe doet. Omdat deze echter altijd in de geselecteerde modellen zat en het bijna verworpen was nemen we het toch mee in ons model. Omdat normaliteit van de regressoren de kwaliteit van een model sterk beinvloeden gebruiken we dezelfde transformaties als in sectie 2. Voor de interpretatie hanteren we echter een model zonder transformaties om ons meer zinvol te kunnen uitspreken over het model.

3.2 Model evaluatie

De Wald test verwerpt het niet significant zijn van de enkele variabelen allemaal op niveau 0.05 met voldoende marge. De behaalde deviance is 6305.7 en de AIC 6327.7. Dit is een daling van ongeveer 1000 tegenover het model zonder transformaties. De likelihood ratio test verwerpt ook voor alle regressoren dat ze niet significant zouden zijn op niveau 0.05 met voldoende marge. De deviance residuals (figuur 8) vertonen niets zorgbarends, behalve dat er een opslitsing te zien is tussen de steden, mede veroorzaakt omdat ze aan elkaar geplakt zijn en dus gegroepeerd in de data. De goodness of fit test verwerpt ook sterk op significantieniveau 0.05. Qua problemen is er de duidelijke opsplitsing van de deviance residuals, maar de oorzaak is bekend en geeft geen verdere problemen. Doordat de residuals niet normaal verdeeld zijn bij logistische regressie kan er ook niet sterk gecontroleerd worden op problemen met normaliteit of modelveronderstelling om problemen te vinden.



Figuur 8: Deviance residuals.

3.3 Model interpretatie

Het model met dezelfde variabelen maar ongetransformeerd
kan als volgt geïnterpreteerd worden:
Iemand in Brussel heeft 54 procent meer odds dat het vol zit dan in Antwerpen. Iemand in Gent 73% meer kans dat het vol zit dan in Antwerpen. Iemand die een privé kamer zoekt heeft 22.5 procent meer kans dat het vol zit dan bij een volledig huis of appartement. Iemand die een gedeelde kamer zoekt heeft 16.2 procent minder kans dat het vol zit dan bij een volledig huis of appartement. Je hebt meer kans dat het verblijf permanent verhuurd is als:

- prijs - (1.2)
- laatste review + (0.3)
- aantal reviews - (0.7)
- reviews maand - (22.1)
- minimum nachten - (3)
- zoekertjes uitbater - (4)

waarbij -/+ betekent de kans stijgt als het attribuut daalt/stijgt. Het cijfer tussen de haakjes is het aantal odds dat gestegen of gedaald wordt bij het veranderen van 1 van het attribuut. Als het langer geleden is dat de laatste reviews werd geschreven heb je een grotere kans dat het verblijf permanent verhuurd is. Als de prijs, het aantal reviews in totaal en per maand, het minimum aantal nachten en het aantal zoekertjes per uitbater lager is, heb je een grotere kans dat het verblijf permanent verhuurd is. Dit strookt met wat je zou denken dat er in de realiteit gebeurt, behalve dan het minimum aantal nachten. Dit is waarschijnlijk het gevolg van de uitbaters die dit op 0 zetten als default als hun verblijf permanent verhuurd is.

Besluit

Afsluitende tekst.