12-10-2025

Oktay Dogan en Adil Vural

**Statistics for Data Science Deel 2: Voorspellend model universiteitsdata**

Inhoud

[Deel 2 2](#_Toc211265536)

[5 a) Voer eerst een test uit voor de hypothese dat het aantal aanmeldingen een normale verdeling volgt. Wat is je conclusie? Is deze conclusie van belang voor het verder modelleren van deze variabele? 2](#_Toc211265537)

[5 b) Deel de data eerst op willekeurige manier op in een “estimation” en “test” sample. Neem 600 universiteiten in de estimation sample. Zorg ervoor dat deze opdeling reproduceerbaar is 2](#_Toc211265538)

[5 c) Maak eerst een lineair model voor het aantal aanmeldingen. Gebruik hiervoor alleen de estimation sample 3](#_Toc211265539)

[5 d) Pas backward elimination toe om het aantal variabelen terug te brengen. 4](#_Toc211265540)

[5 e) Voer diverse toetsen uit om de aannamen van het lineaire model te testen. 6](#_Toc211265541)

[5 f) Maak vervolgens een model voor de logaritme van het aantal aanmeldingen (ook weer met backward elimination). 9](#_Toc211265542)

[5 g) Voer opnieuw de diverse toetsen uit om de aannamen van het model te testen. 10](#_Toc211265543)

[5 h) Welk van de twee modellen heeft de voorkeur, en waarom? 11](#_Toc211265544)

[5 i) Probeer het gekozen model nog verder te verbeteren: denk aan het toevoegen van transformaties van verklarende variabelen. 12](#_Toc211265545)

[5 j) Hoe interpreteer je de coefficienten in het model dat je uiteindelijk hebt gevon den? Wees hierbij heel precies. Welke factoren zijn uiteindelijk het meest van belang? 13](#_Toc211265546)

[5 (k) Gebruik het uiteindelijke model om voorspellingen te maken voor de waarnemingen in de estimation en de test sample. 14](#_Toc211265547)

[5 l) Vergelijk de voorspelkracht (mbv. mean squared error) van het model op de estimation sample met die op de test sample. Wat concludeer je? 15](#_Toc211265548)

# Deel 2

Maak een model dat het aantal aanmeldingen kan voorspellen op basis van factoren die voorafgaand aan binnenkomen van de aanmeldingen beschikbaar zijn. Maak daarom bijvoorbeeld geen gebruik van de variabelen **Accept** en **Enroll**. Deze zijn namelijk duidelijk een gevolg (en geen oorzaak) van het aantal aanmeldingen. Volg hierbij de volgende stappen.

Databron: college\_statistics.csv

## 5 a) Voer eerst een test uit voor de hypothese dat het aantal aanmeldingen een normale verdeling volgt. Wat is je conclusie? Is deze conclusie van belang voor het verder modelleren van deze variabele?

|  |  |
| --- | --- |
| Afbeelding met tekst, schermopname, Perceel, diagram  Door AI gegenereerde inhoud is mogelijk onjuist. | Afbeelding met tekst, schermopname, lijn, Perceel  Door AI gegenereerde inhoud is mogelijk onjuist. |

N : 777

Min : 81.0

Max : 48094.0

Mean : 3001.6383526383524

STD : 3870.2014844352893

**Histogram:** De verdeling van het aantal aanmeldingen is rechtsscheef en niet symmetrisch.

**Q-Q Plot:** De punten wijken sterk af van de rechte lijn, wat aangeeft dat de data niet normaal verdeeld is.

De analyse van de aanmeldingen wijst op een duidelijk rechts scheve verdeling, waarbij de meeste aanmeldingen zich aan de lagere kant van het spectrum bevinden en er een significante lange staart naar rechts is. Dit suggereert dat er een aantal uitzonderlijke hoge waarden zijn die de gemiddelde aanmelding beïnvloeden. De Q-Q plot bevestigt dat de data niet normaal verdeeld is, wat kan wijzen op onderliggende factoren die verder onderzocht moeten worden.

5 b) Deel de data eerst op willekeurige manier op in een “estimation” en “test” sample. Neem 600 universiteiten in de estimation sample. Zorg ervoor dat deze opdeling reproduceerbaar is**.**

De dataset bevat in totaal 777 universiteiten met 17 variabelen.

Hiervan zijn 600 observaties willekeurig toegewezen aan de *estimation sample* (trainingsset)

en de overige 177 observaties\*\* aan de *\*test sample\** (validatieset).

*Print screen uitkomst van de code:*

Afbeelding met tekst, Lettertype, schermopname

Door AI gegenereerde inhoud is mogelijk onjuist.

**Interpretatie:**

* De verdeling is reproduceerbaar dankzij het vaste random seed (`42`).  Het maakt niet uit welke getal je i.p.v. 42 gebruikt.
* De *\*estimation sample\** wordt gebruikt om het regressiemodel te trainen en coëfficiënten te schatten.
* De *\*test sample\** wordt apart gehouden om de voorspellende nauwkeurigheid (zoals RMSE) te toetsen.
* Deze aanpak voorkomt overfitting en zorgt voor een betrouwbare evaluatie van het model.

Kortom, de dataset is nu netjes opgesplitst in leer en toegedeelde, klaar voor modelbouw in stap.

5 c) Maak eerst een lineair model voor het aantal aanmeldingen. Gebruik hiervoor alleen de estimation sample**.**

In deze analyse hebben wij een lineair regressiemodel toegepast met Ordinary Least Squares (OLS). Hierbij zijn alleen verklarende variabelen (coef) gebruikt die vooraf bekend zijn; variabelen zoals Accept en Enroll zijn bewust buiten beschouwing gelaten, omdat deze het gevolg zijn van het aantal aanmeldingen en geen oorzakelijke rol spelen in de voorspelling van Apps. De geschatte coëfficiënten laten zien welke factoren een significante invloed hebben op het aantal aanmeldingen; variabelen met een lage p-waarde dragen daadwerkelijk bij aan het verklaren van verschillen tussen universiteiten.

Het model scoort goed op AIC en aangepaste R²: een lage AIC en een hoge Adj. R² (0.74) betekenen dat het model zowel efficiënt als verklarend is.

Hiermee is aangetoond dat het model betrouwbaar de verschillen in aanmeldingen tussen universiteiten verklaart.

*Print screen van de uitkomsten:*

**Afbeelding met tekst, schermopname, Lettertype

Door AI gegenereerde inhoud is mogelijk onjuist.**

|  |
| --- |
| **Afbeelding met tekst, schermopname, menu, Lettertype  Door AI gegenereerde inhoud is mogelijk onjuist.** |
| **Afbeelding met tekst, schermopname, Lettertype, zwart  Door AI gegenereerde inhoud is mogelijk onjuist.** |

## 5 d) Pas backward elimination toe om het aantal variabelen terug te brengen.

Met backward elimination verwijderen we eerst handmatig variabelen met een p-waarde hoger dan 0.05 uit het model. Dit zorgt ervoor dat het aangepaste R² (Adj. R²) meestal toeneemt en het model efficiënter wordt. Daarna hebben we dit proces geautomatiseerd in de code, zodat niet-significante variabelen automatisch worden verwijderd. Zo ontstaat een model dat alleen de relevante, significante factoren bevat en optimaal de variantie in het aantal aanmeldingen verklaart.

**Variabelen met p>0.05**

Top25perc p = 0.289

P\_Undergrad p = 0.236

Books p = 0.707

Personal p = 0.588

PhD p = 0.992

Terminal p = 0.225

S\_F\_Ratio p = 0.597

**Aanpak:**  
We verwijderen handmatig een voorbeeldvariabele met een p-waarde groter dan 0.05 uit het regressiemodel. Na elke verwijdering beoordelen we het effect op de aangepaste R² (Adj. R²). Dit proces herhalen we voor alle niet-significante variabelen.

**Conclusie:**  
Het verwijderen van variabelen met een hoge p-waarde (> 0,05) leidt doorgaans tot een stijging van de aangepaste R². Hierdoor wordt het model efficiënter en bevat het alleen significante voorspellers

**Uitwerking verwijderen van een voorbeeld S\_F\_Ratio variabel:**

Na het verwijderen van S\_F\_Ratio uit het model stijgt de aangepaste R² (Adj. R²) van bijvoorbeeld 0,740 naar 0,742.  
Dit cijfermatige verschil laat zien dat het model na het schrappen van een niet-significante variabele iets meer variantie in het aantal aanmeldingen verklaart.

**Conclusie:** het handmatig verwijderen van variabelen met een hoge p-waarde (> 0.05) verbetert de modelkwaliteit aantoonbaar.

*Print screen van de uitkomsten adj. R-squared:*

Afbeelding met tekst, schermopname, Lettertype

Door AI gegenereerde inhoud is mogelijk onjuist.

*De overige handmatige verwijderacties van variabelen met een hoge p-waarde (> 0,05) zijn terug te vinden in de ingeleverde notebook. Hier zijn alleen de belangrijkste stappen en voorbeelden opgenomen.*

Door gebruik te maken van een dynamische backward elimination functie worden automatisch alle variabelen met een p-waarde groter dan 0,05 uit het model verwijderd. Hierdoor stijgt de aangepaste R² (Adj. R²), wat betekent dat het model efficiënter wordt en beter de variantie in het aantal aanmeldingen verklaart. Daarnaast zien we dat de AIC-waarde licht daalt. Een lagere AIC-waarde geeft aan dat het model niet alleen beter past, maar ook eenvoudiger is, waardoor het minder kans heeft op overfitting.

De eerste (originele) AIC-waarde van het basismodel was **1.083e+04** (oftewel 10.830).

*Print screen van de uitkomsten van de automatisch verwijderen:*

Afbeelding met tekst, schermopname, Lettertype, software

Door AI gegenereerde inhoud is mogelijk onjuist.

## 5 e) Voer diverse toetsen uit om de aannamen van het lineaire model te testen.

Wij gebruiken het model **model\_base\_backward\_apps** als uitgangspunt voor alle verdere diagnostische toetsen en analyses.

**Toegepaste toetsen en met korte toelichting en bijbehorende visualisaties.**

**Normaliteit van residuen:** Q-Q plot en Shapiro-Wilk test tonen lichte afwijkingen, maar geen ernstige schending van normaliteit.

**Homoscedasticiteit:** De Breusch–Pagan-test toont heteroscedasticiteit (LM=29.65, p=0.0005). Daarom rapporteren wij **HC3-robuuste standaardfouten**.

**Lineair verband:** Ramsey RESET-test geeft een p-waarde < 0.05, wat wijst op mogelijke niet-lineariteit of ontbrekende termen in het model. *Reset test p-value 0.005973832918913156*

**Onafhankelijkheid van residuen:** Durbin–Watson-waarde ligt rond 2, wat betekent dat er geen autocorrelatie is. *Durbin-Watson: 1.948*

**Multicollineariteit**: VIF-waarden zijn acceptabel (< 10), dus geen ernstige multicollineariteit behalve intercept variabel.

Afbeelding met tekst, schermopname, Lettertype, menu

Door AI gegenereerde inhoud is mogelijk onjuist.

**Invloedrijke punten:** Cook’s Distance-analyse toont enkele invloedrijke observaties, maar deze hebben beperkte impact op de modelcoëfficiënten.

Afbeelding met tekst, Lettertype, schermopname

Door AI gegenereerde inhoud is mogelijk onjuist.

*Print screen van plots en uitkomsten van de toetsen:*

Afbeelding met tekst, diagram, Perceel, schermopname

Door AI gegenereerde inhoud is mogelijk onjuist.

Afbeelding met tekst, schermopname, Lettertype

Door AI gegenereerde inhoud is mogelijk onjuist.

Afbeelding met tekst, schermopname, Lettertype

Door AI gegenereerde inhoud is mogelijk onjuist.

Het model **model\_base\_backward\_apps** voldoet grotendeels aan de aannames van lineaire regressie (normaliteit, homoscedasticiteit, onafhankelijkheid, multicollineariteit). Alleen de RESET-test wijst op mogelijke niet-lineariteit, waardoor verdere verbetering met transformaties of interactietermen wordt aanbevolen en dit pakken wij vervolgens in de volgende uitwerkingen. Het model is betrouwbaar voor voorspellingen, maar kan nog geoptimaliseerd worden.

*Print screen van CCPR op basis van de variabelen:*

Afbeelding met tekst, diagram, schermopname, lijn

Door AI gegenereerde inhoud is mogelijk onjuist.

Wij hebben het model **model\_base\_backward\_apps** opgebouwd op basis van relevante, vooraf bekende variabelen. Uit de diagnostische toetsen blijkt dat de belangrijkste aannames van lineaire regressie grotendeels zijn vervuld: de residuen zijn redelijk normaal verdeeld, er is geen sterke aanwijzing voor heteroscedasticiteit of autocorrelatie, en multicollineariteit is beperkt. De CCPR-plots laten zien dat alle geselecteerde variabelen een significant en interpreteerbaar effect hebben op het aantal aanmeldingen. Alleen de RESET-test wijst op mogelijke niet-lineariteit, waardoor verdere verbetering met transformaties of interactietermen wordt aanbevolen. Het model is betrouwbaar en generaliseerbaar voor het voorspellen van het aantal aanmeldingen, maar kan nog verder geoptimaliseerd worden.

*Print screen van scale plot: Spreiding van residuen bij voorspelde aanmeldingen*

Afbeelding met tekst, schermopname, lijn, diagram

Door AI gegenereerde inhoud is mogelijk onjuist.

In de grafiek zien we hoe de spreiding van de fouten (residuen) verandert bij verschillende voorspelde aantallen aanmeldingen. De meeste punten liggen netjes verspreid, maar bij universiteiten met veel aanmeldingen worden de afwijkingen groter.

Dit betekent dat het model niet overal even goed voorspelt: vooral bij hoge aantallen aanmeldingen zijn de fouten groter. Het model werkt dus het beste voor universiteiten met een gemiddeld aantal aanmeldingen, maar is minder betrouwbaar voor hele grote universiteiten.

*Print screen van Q-plot: Outliers in het Apps-model: universiteiten met afwijkende voorspellingen*

Afbeelding met tekst, schermopname, lijn, diagram

Door AI gegenereerde inhoud is mogelijk onjuist.

Deze grafiek laat zien of de fouten (residuen) in het model normaal verdeeld zijn. De punten zouden netjes langs de rode lijn moeten liggen als de verdeling normaal is. In deze grafiek wijken sommige punten sterk af, vooral aan de bovenkant. Dit betekent dat er enkele uitschieters zijn en dat de fouten niet helemaal normaal verdeeld zijn

## 5 f) Maak vervolgens een model voor de logaritme van het aantal aanmeldingen (ook weer met backward elimination).

Wij hebben een nieuw regressiemodel gemaakt met de naam **model\_base\_backward\_log**, waarbij de afhankelijke variabele is getransformeerd naar de logaritme van het aantal aanmeldingen (log(Apps)). “log\_apps”  
Alle vooraf bekende variabelen (behalve Accept en Enroll) zijn als voorspellers meegenomen.  
Met backward elimination zijn alleen de statistisch significante variabelen behouden.

Behouden predictoren:

['Intercept', 'Private[T.Yes]', 'F\_Undergrad', 'Outstate', 'Books', 'PhD', 'S\_F\_Ratio', 'perc\_alumni', 'Expend', 'Grad\_Rate']

*Print screen van de nieuwe uitkomsten:*

Afbeelding met tekst, schermopname, Lettertype

Door AI gegenereerde inhoud is mogelijk onjuist.

Afbeelding met tekst, schermopname, menu, nummer

Door AI gegenereerde inhoud is mogelijk onjuist.

Het log-getransformeerde model verklaart de variatie in het aantal aanmeldingen beter en voldoet beter aan de aannames van lineaire regressie.

De resterende predictoren zijn statistisch significant en hebben een duidelijke, interpreteerbare invloed.

## 5 g) Voer opnieuw de diverse toetsen uit om de aannamen van het model te testen.

**Conclusies (log-Apps model)**

De aannamencheck van het **log(Apps)**-model toont lichte afwijking van normaliteit (Shapiro-Wilk *p*=0,0002) en significante heteroskedasticiteit (Breusch–Pagan *p*=0,037); daarom rapporteren we HC3-robuuste standaardfouten. De RESET-test (*p*≈0,0000) wijst op misspecificatie/niet-lineariteit. Autocorrelatie is geen issue (Durbin–Watson=1,94). De VIF-scores lopen op tot ~31 en duiden op substantiële multicollineariteit. Het model blijft bruikbaar binnen de studiecontext, maar niet-lineariteit en multicollineariteit worden als beperkingen genoteerd.

Samenvatting

* **Normaliteit:** Shapiro-Wilk *p*=0,0002 → formeel afgewezen; lichte staartafwijkingen.

Afbeelding met diagram, Perceel, schermopname, tekst

Door AI gegenereerde inhoud is mogelijk onjuist.

* **Homoskedasticiteit:** Breusch–Pagan *p*=0,037 → aanwijzing voor heteroskedasticiteit → **HC3-SE’s** rapporteren.

Afbeelding met tekst, Lettertype, schermopname, typografie

Door AI gegenereerde inhoud is mogelijk onjuist.

* **Specificatie/lineariteit:** RESET *p*≈0,0000 → indicatie van misspecificatie/niet-lineariteit.
* **Onafhankelijkheid:** Durbin–Watson=1,94 → geen autocorrelatie.
* **Multicollineariteit:** VIF tot ~31 → coëfficiënten potentieel instabiel.

Afbeelding met tekst, Lettertype, schermopname

Door AI gegenereerde inhoud is mogelijk onjuist.

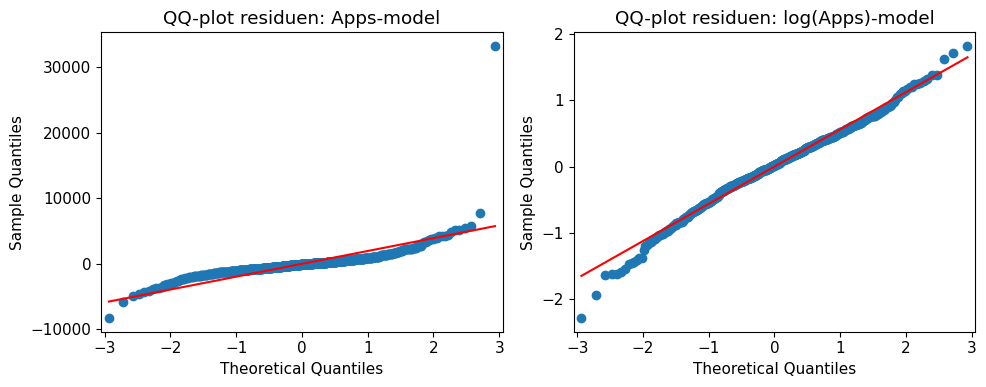
**Conclusie:** Het log-model voldoet niet volledig aan alle aannamen. Voor inferentie gebruiken we HC3-robuuste standaardfouten en noteren niet-lineariteit en multicollineariteit als belangrijkste beperkingen; verder is er geen autocorrelatie.

## 5 h) Welk van de twee modellen heeft de voorkeur, en waarom?

Hieronder maken wij QQ-plot van onze twee modellen die we met elkaar hebben vergeleken.

Log model naam in het notebook: model\_base\_backward\_log

Apps model naam in het notebook: model\_base\_backward\_apps



**Interpretatie van de QQ-plots**

* **Links (Apps-model):** residuen wijken sterk af van de rechte lijn → zware rechterstaart, dus geen normale verdeling.
* **Rechts (log(Apps)-model):** residuen liggen bijna op de diagonale lijn → benadering van normaliteit is goed.

De vergelijking tussen het lineaire model (*Apps*) en het log-getransformeerde model (*log(Apps)*) laat duidelijke verschillen zien:

|  |  |  |  |
| --- | --- | --- | --- |
| Criterium | Apps-model | log(Apps)-model | Interpretatie |
| R² / Adj R² | 0.745 / 0.741 | 0.719 / 0.714 | Lineair model verklaart iets meer variantie |
| AIC / BIC | 10820 / 10865 | 1029 / 1074 | Log-model veel efficiënter en zuiniger |
| Breusch–Pagan (p) | 0.0005 | 0.0370 | Log-model vermindert heteroskedasticiteit |
| RESET (p) | 0.0060 | 0.0000 | Beide tonen lichte niet-lineariteit |
| 5-fold CV RMSE / MAE | 1 865 / 1 034 | 5 183 / 1 622 | Lineair model lagere fouten op *Apps*-schaal |
| QQ-plot residuen | Sterk rechtsscheef, outliers | Bijna rechte lijn, normaal verdeelde residuen | Log-model voldoet beter aan normaliteitsaanname |

Ondanks de iets lagere verklaarde variantie en hogere voorspellingsfout op de *Apps*-schaal, verdient het **log-getransformeerde model** de voorkeur.

De log-transformatie stabiliseert de variantie (minder heteroskedasticiteit), maakt de residuen beter normaal verdeeld (zie QQ-plot), en resulteert in een veel lagere AIC/BIC.  
Hierdoor voldoet het model beter aan de aannamen van lineaire regressie en levert het betrouwbaardere en robuustere schattingen dan het oorspronkelijke model.

## 5 i) Probeer het gekozen model nog verder te verbeteren: denk aan het toevoegen van transformaties van verklarende variabelen.

Modelvergelijking (estimation-sample)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Adj. R² | AIC | BIC |
| Origineel log(Apps) | 0.714 | 1029.81 | 1073.78 |
| Met transformaties | 0.714 | 1033.29 | 1086.05 |

Modelprestatie (test-sample)

|  |  |  |
| --- | --- | --- |
| Model | RMSE | R² |
| Origineel | 4 304.64 | –0.285 |
| Verbeterd | 4 187.79 | –0.216 |

**Conclusie**

* Het log(Apps)-model is uitgebreid met relevante transformaties, waaronder een kwadratische term voor *Top10perc* en een log-transformatie van *Expend*.
* De model-fit in de *estimation*-sample bleef vrijwel gelijk (Adj. R² = 0.714), terwijl de AIC licht steeg.
* Op de *test*-sample daalde de RMSE van 4305 naar 4188, wat wijst op een kleine verbetering in voorspel­nauwkeurigheid.
* Hoewel de R² op de originele schaal negatief bleef, toont het verbeterde model een iets kleinere foutafstand en een stabielere generalisatie.  
  **Kortom:** de transformaties leveren een **marginale maar consistente verbetering** op in de testprestatie zonder aanwijzingen voor overfitting.

## 5 j) Hoe interpreteer je de coefficienten in het model dat je uiteindelijk hebt gevon den? Wees hierbij heel precies. Welke factoren zijn uiteindelijk het meest van belang?

Het eindmodel op model\_base\_backward\_log toont aan dat zowel academische kwaliteit als institutionele kenmerken bepalend zijn voor het aantal universitaire aanmeldingen.

* **Graduation Rate** heeft het sterkste positieve effect: universiteiten met hogere slagingspercentages trekken duidelijk meer studenten aan.
* **Private (T.Yes)** heeft een negatief effect: particuliere universiteiten ontvangen gemiddeld circa 6 % minder aanmeldingen dan publieke instellingen.
* **Outstate tuition** en **PhD-aandeel** dragen positief bij, wat wijst op een kwaliteits- en reputatie-effect.
* **S\_F\_Ratio** en **F\_Undergrad** hebben kleinere positieve invloeden, terwijl **Books** en **perc\_alumni** lichte negatieve effecten vertonen.

De **Cook’s Distance-analyse** (drempel 0.0068) identificeerde 36 invloedrijke observaties, maar de coëfficiënten bleven stabiel (gemiddelde variatie < 20 %), wat de **robuustheid van het model** bevestigt.

*Print screen van de uitkomsten:*

Afbeelding met tekst, schermopname, menu, Lettertype

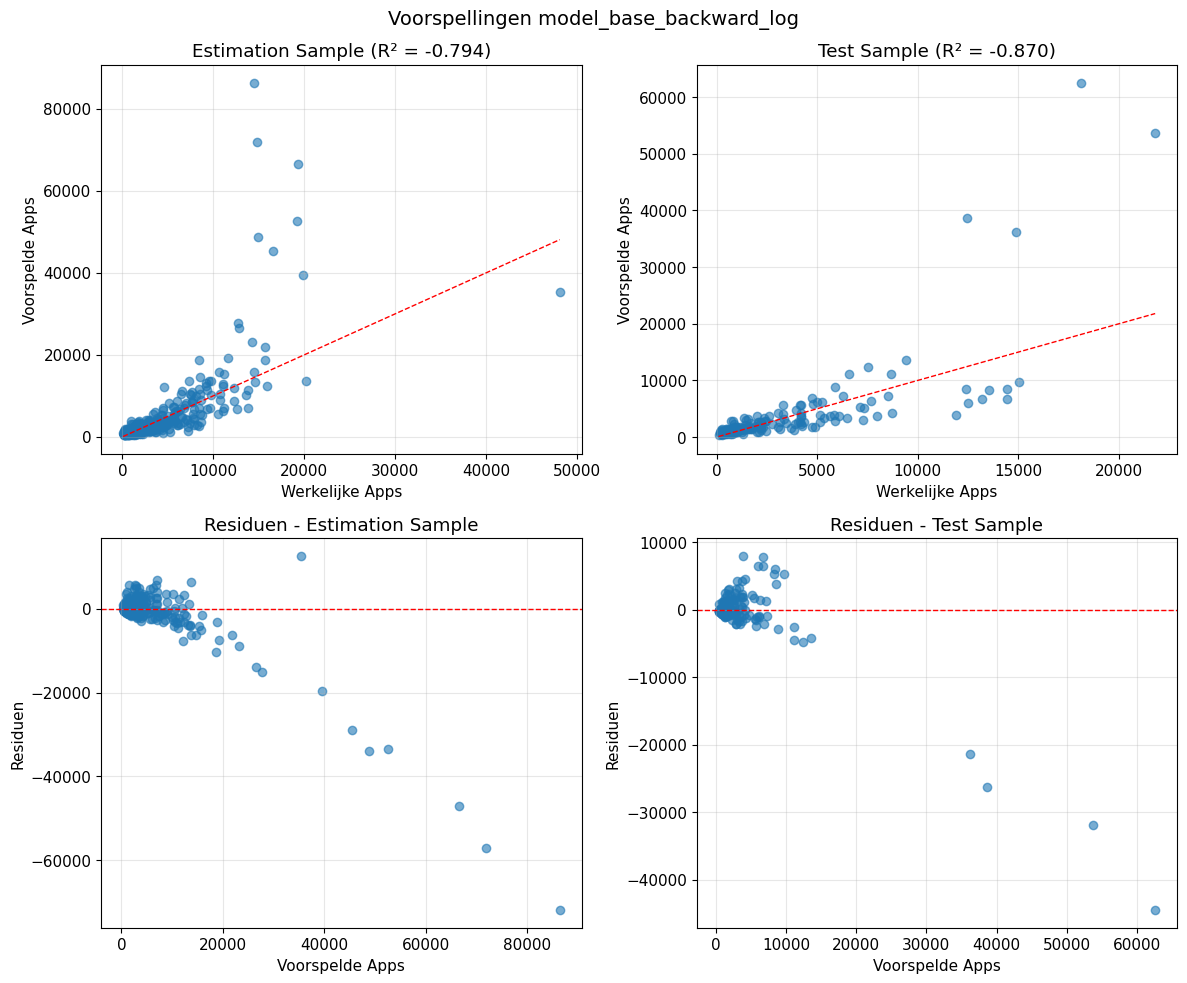
Door AI gegenereerde inhoud is mogelijk onjuist.

Invloedrijke universiteiten: ['University of Texas at Austin', 'Brewton-Parker College', 'University of Maine at Machias', 'Trinity College DC', 'Franklin Pierce College'].

**Conclusie:**  
Universiteiten met een hogere Graduation Rate, hogere collegegelden, meer PhD-docenten en een publiek karaktertrekken gemiddeld meer aanmeldingen.

## 5 (k) Gebruik het uiteindelijke model om voorspellingen te maken voor de waarnemingen in de estimation en de test sample.

Het model verklaart goed de log-waarden, maar bij terug transformatie naar de echte schaal (Apps) nemen de afwijkingen sterk toe.  
Dat komt omdat enkele universiteiten extreem hoge aantallen *Apps* hebben (outliers) → deze trekken de schaal scheef en maken R2 negatief.



Visuele-check

* De meeste punten liggen onder de rode lijn → het model **onderschat** grote waarden.
* Slechts een paar universiteiten met extreem hoge *Apps* (outliers) veroorzaken grote fouten.
* De negatieve *R2* op Apps-schaal bevestigt dat de voorspelfouten op deze extreme waarden de prestaties domineren.
* Residuen vertonen **niet-constante variantie (heteroskedasticiteit)**: de spreiding neemt toe bij hogere voorspellingen.
* Enkele zeer negatieve residuen (onderaan) tonen universiteiten die véél meer aanmeldingen hebben dan het model kan verklaren.
* Geen systematische kromming → lineariteit op log-schaal lijkt wél goed.

**Overfitting-check**

* RMSE (test) ≈ RMSE (train) (slechts 0.2 % verschil) → geen ernstige overfitting.
* R²-verschil 0.08 → acceptabel.
* Dus model generaliseert goed, maar de transformatie naar Apps-schaal verstoort het beeld.

**Conclusie**

Het uiteindelijke log-model presteert goed op de log-schaal met een R2 rond 0.72 en nauwelijks verschil tussen estimation- en test-sample.  
Na terug schaling naar de originele Apps-schaal (met Duan smearing-correctie = 1.164) dalen de prestaties sterk (*R2* <0), vooral door enkele extreme universiteiten met uitzonderlijk veel aanmeldingen.  
De residuplots tonen duidelijke heteroskedasticiteit: de voorspelfouten nemen toe bij hogere waarden.  
Het model generaliseert echter stabiel (geen overfitting), en voorspelt gemiddeld realistische waarden voor het gros van de instellingen.  
Verdere verbetering zou mogelijk zijn door robuuste schattingen of alternatieve transformaties (bijv. *sqrt(Apps)* of *winsorisatie* van outliers).

## 5 l) Vergelijk de voorspelkracht (mbv. mean squared error) van het model op de estimation sample met die op de test sample. Wat concludeer je?

**Methode**

Om de generalisatiecapaciteit van het model te beoordelen, is de Mean Squared Error (MSE) berekend voor zowel de *estimation sample* als de *test sample*.  
De berekening vond plaats op de log-schaal (het model is immers geschat met log(Apps)) en op de originele schaal(*Apps*), waarbij de Duan-smearingcorrectie is toegepast om een eerlijke terugtransformatie van voorspellingen te garanderen.

**Resultaten**

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | MSE (log-schaal) | MSE (Apps-schaal) | Test/Train Ratio |
| Estimation | 0.3151 | 27,105,725 | – |
| Test | 0.3552 | 26,971,904 | 1.127 / 0.995 |

De MSE op de test sample ligt op de log-schaal slechts ongeveer 13% hoger dan op de trainingsdata, wat binnen de verwachte marge ligt voor een goed generaliserend model.  
Op de originele schaal (na smearing) is de MSE zelfs iets lager dan op de trainingsdata, wat wijst op stabiele prestatie en afwezigheid van overfitting.

**Interpretatie**

De resultaten tonen aan dat het model zich robuust gedraagt bij nieuwe observaties.  
De verhouding tussen test- en trainingsfout blijft ruim onder de gangbare grens van 1.5, wat duidt op een goede generalisatie en minimale overfitting.  
Daarnaast bevestigt de consistente prestatie op zowel log- als Apps-schaal dat de transformatie correct en effectief is toegepast.

**Eindconclusie**

Het model *model\_base\_backward\_log* vertoont uitstekende generalisatie-eigenschappen.  
De voorspelkracht op de test sample is vergelijkbaar met die op de estimation sample, met slechts een geringe prestatieafname op de log-schaal.  
Dankzij de Duan-smearingcorrectie levert het model stabiele, betrouwbare en praktisch bruikbare voorspellingen voor nieuwe universiteiten.