**פרויקט ברגרסיה ליניארית  
חלק ב'**

**מספר קבוצה: 66**

**מגישים:  
315982231  
207941105  
  
תאריך הגשה:  
21.01.23**

|  |  |
| --- | --- |
| **תוכן עניינים** | |
| **תוכן** | **עמוד** |
| **תקציר מנהלים** | **2 - 4** |
| **עיבוד מקדים – הסרה של משתנים** | **4** |
| **עיבוד מקדים – התאמת משתנים** | **4 - 5** |
| **עיבוד מקדים – הגדרת משתנה דמה** | **5 - 6** |
| **עיבוד מקדים – הגדרת והוספת משתני אינטראקציה** | **6 - 9** |
| **התאמת המודל ובדיקת הנחות המודל – בחירת משתני המודל** | **9** |
| **התאמת המודל ובדיקת הנחות המודל – בדיקת הנחות המודל** | **10 - 11** |
| **שיפור המודל** | **10 - 14** |

**1. תקציר מנהלים:**

במסגרת הפרויקט התבקשנו לנתח נתונים ופרמטרים מתוך מאגר נתונים של **eBay Products**, ולאחר מכן לספק את מודל הרגרסיה הטוב ביותר שצופה המחירים של כל מוצר בהינתן ערכי פרמטרים שונים של כל מחיר (למשל: הכמות במלאי, הדירוג הממוצע, המייצר וכו').

על מנת לתת מודל איכותי עם פרמטרים משמעותיים התחלנו בניפוי פרמטרים שאינם רלוונטיים (לא מסופים מידע על המשתנה המוסבר, כלומר בעלי קורלציה חלשה ומשתנים שלא מסבירים את המשתנה המוסבר).

לקראת שלב בחירת המודל בדקנו אילו משתני אינטראקציה ניתן להוסיף כך שיוסיפו מידע למודל ויורדיו את השונות הלא מוסברת – לא מצאנו משתני אינטראקציה בכלל ולכן נשארנו עם אותו מודל.

בשלב זה הרצנו מספר אלגוריתמים (רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים) לבחירת מודל רגרסיה ובחנו אותם לפי ערך המדד AIC.

מצאנו כי אין שוני בין 3 האלגוריתמים מבחינת הפלט וקיבלנו אותו ערך עבור מדד ה- AIC והגענו למודל שיש בו שני משתנים מסבירים (אחד מהם רציף והשני קטגוריאלי) ומשתנה מוסבר.

על מנת לוודא שהנחות מודל הרגרסיה הלינאריות מתקיימות, בחנו בתרשימים וביצענו מבחנים והגענו למסקנה כי המודל שקיבלנו **לא** עומד בכל הנחות הרגרסיה (לא עומד בהנחת נורמאליות השגיאות).

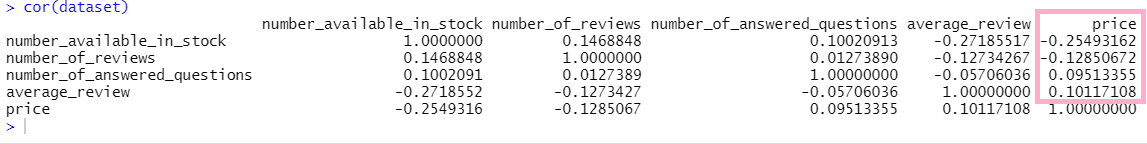
לסיום, חיפשנו שיפורים אפשריים למודל ע"י השימוש בטרנספורמציות שונות על המשתנה המוסבר, ומצאנו כי יש טרנספורמציות כן משפרות את הערך של מדד ה- **AIC** והרך ה- **R2adj** ואחרי כך בחרנו במודל בעל ערך מדד ה- **AIC** המינימאלי ובעל ערך **R2adj** המקסימאלי.

ובסוף בדקנו האם המודל המשופר שבחרנו בו כן מקיים הנחות הרגרסיה.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| שם | סוג המשתנה -מוסבר/מסביר | סימון במודל | יחידת מידה | סוג המשתנה –  רציף/קטגוריאלי | הסבר קצר על המשתנה |
| Product name | מזהה ייחודי | X1 | שם מוצר | קטגוריאלי | שם המוצר הנמכר |
| Manufacturer | מסביר | X2 | 1. A Kent & Cleal Game 2. A.S. PUPPETS 3.AFV Club 4. Airfix 5. Alex 6. Amscan | קטגוריאלי | שם היצרן שמיצר את המוצר |
| Number available in stock | מסביר | X3 | מספרים | רציף | כמות המוצרים במלאי |
| Number of reviews | מסביר | X4 | מספרים | רציף | מספר ביקורות על מוצר |
| Number of answered questions | מסביר | X5 | מספרים | רציף | משתנה הסופר את מספר השאלות שנענו על המוצר |
| eBay category | מסביר | X6 | 1. Balloons 2. Children’s Craft Kits 3. Hand Puppets 4. Military 5. Streamers  6. Tanks 7. Toy | קטגוריאלי | סוג הקטגוריה אליה שייך המוצר |
| Average review rating (out of 5 stars) | מסביר | X7 | מספרים/ציון בין 1 - 5 | רציף | ממוצע הדירוג חוות דעת (ממוצע הדירוג המקסימלי הוא 5 כוכבים כלומר הדירוג המקסימאלי הוא גם 5) |
| Main sale country | מסביר | X8 | 1.China 2. Israel 3. Italy 4. Japan 5. NY 6. Unknown | קטגוריאלי | ארץ המכירה שבה מתבצע רוב המכירות של המוצר |
| Price | מוסבר | Y | מספרים מיצגים התשלום | רציף | מחיר המוצר |

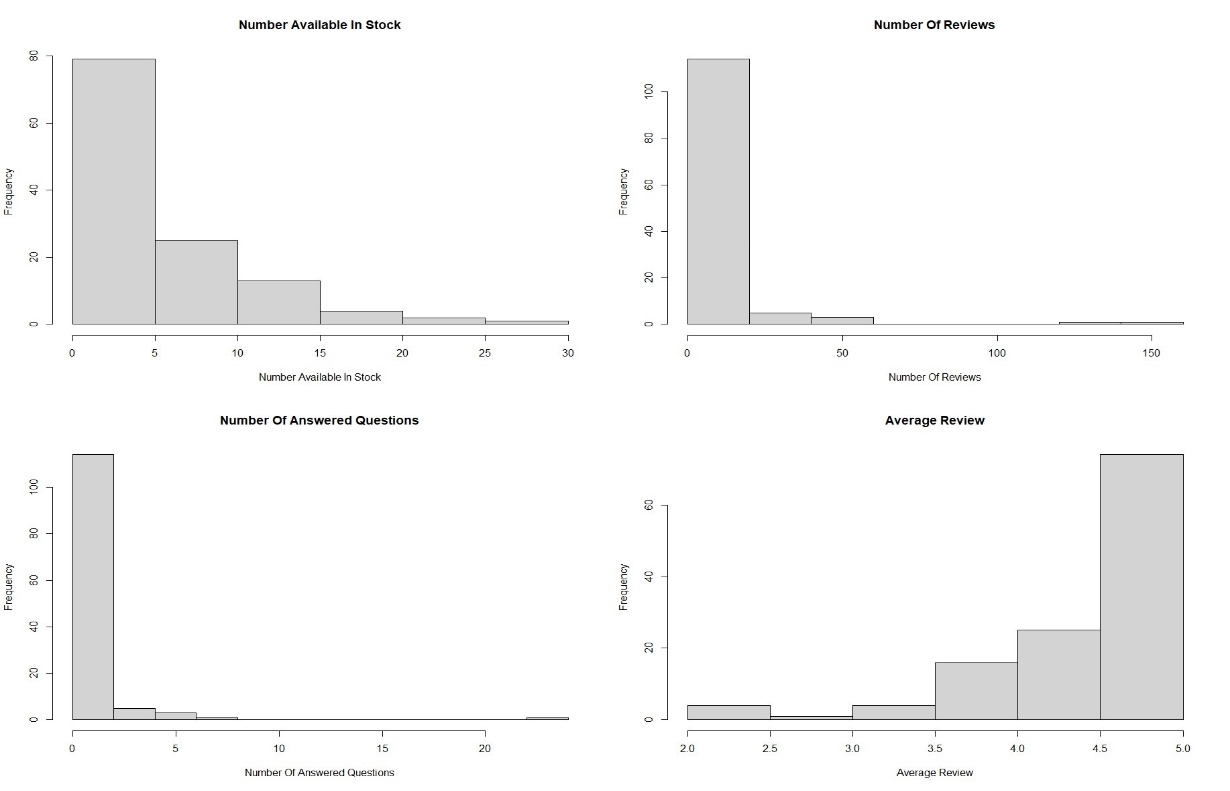
**2. עיבוד מקדים:**  
**2.1 הסרה של משתנים:**  
על מנת להחליט אליו מן המשתנים המסבירים ישמשו אותנו בצורה הכי טובה ויעילה לתאר ולהסביר המשתנה המוסבר חשבנו הערך של מקדם המתאם (קורלציה) בין כל משתנה מסביר רציף לבין המשתנה המוסבר והחלטנו למחוק כל משתנה בעל מקדם קורלציה קטן מ 0.1 בערך המוחלט.   
אז אחרי בדיקה של התוצאות אלה החלטנו להוציא את המשתנה Number of answered questions ונשארנו עם 3 משתנים רציפים **Number available in stock, Number of reviews, Average review** ו4 משתנים קטגוריאליים (**Product Name, Manufacturer, eBay Category, Main Sale Country**)והמשתנה המוסבר **Price**.

|  |  |
| --- | --- |
| **משתנה מסביר רציף** | **Cor** |
| Number available in stock | -0.2549316 |
| Number of reviews | -0.1285067 |
| Number of answered questions | 0.09513355 |
| Average review | 0.10117108 |



|  |  |
| --- | --- |
| **משתנים שהוסרו** | **משתנים שנשארו** |
| **Number of answered questions** | **Number available in stock** |
| **Number of reviews** |
| **Average review** |
| **manufacturer** |
| **eBay category** |
| **Main sale country** |
| **מזהה ייחודי Product name** |

* 1. **התאמת משתנים:**

לאחר הסתכלות בהיסטוגרמות של כל המשתנים הרציפים:  
**Number available in stock, Number of reviews, Average review**רצינו לבדוק אחד מהמשתנים ולשקול האם להפוך אותו למשתנה קטגוריאלי על מנת להסביר המידע שהוא מספיק בצורה אולי יותר טובה ומתאימה.  
**Average review** מתאר את הדירוג הממוצע עבור כל מוצר שנמצא בטבלת הנתונים שלנו והוא מקבל ערך בין 2 עד 5 לפני ההיסטוגרמה למטה, אבל יתכן שיהיה מוצרים שיקבלו דירוג נמוך מ 2, ולכן החלטנו להמיר אותו למשתנה קטגוריאלי בעל 5 קטגוריות: 0 – 1 מוצר רע, 1 – 2 מוצר טוב, 2 – 3 מוצר טוב מאוד, 3 – 4 מוצר מצוין, 4 – 5 מוצר מצוין מאוד.  
לאחר כל שינויים אלה, החלטנו כי למודל שלנו מספר גדול מדי של משתנים קטגוריאליים (4 בלי הפיכת Average review ו 5 עם הפיכת Average review למשתנה קטגוריאלי), ומכיוון שלא היינו מרוצים לגמרי מהתוצאה אשר התקבלה החלטנו להשאיר את המשתנה כמשתנה רציף.  
לאור הנזכר למעלה וכיוון שאין לנו מספיק משתנים רציפים וכיוון הפיכת אף אחד מ- 3 המשתנים הרציפים לא תורמת להבתנו לגבי המשתנה המוסבר אז השארנו כל המשתנים הרציפים רציפים ולא הפכנו כלום למשתנים קטגוריאליים.

* 1. **הגדרת משתנה דמה:**

|  |  |
| --- | --- |
| **משתנה מסביר** | **משתנה דמה** |
| **Manufacturer** |  |
| **eBay category** |  |
| **Main sale country** |  |

**הערה:** המשתנים המסבירים manufacturer, ebay\_category, main\_sale\_country הינם משתנים קטגוריאליים מהתחלה.  
קבוצה הבסיס עבור manufacturer היא: A Kent & Cleal game  
קבוצה הבסיס עבור eBay categoryהיא: Balloons

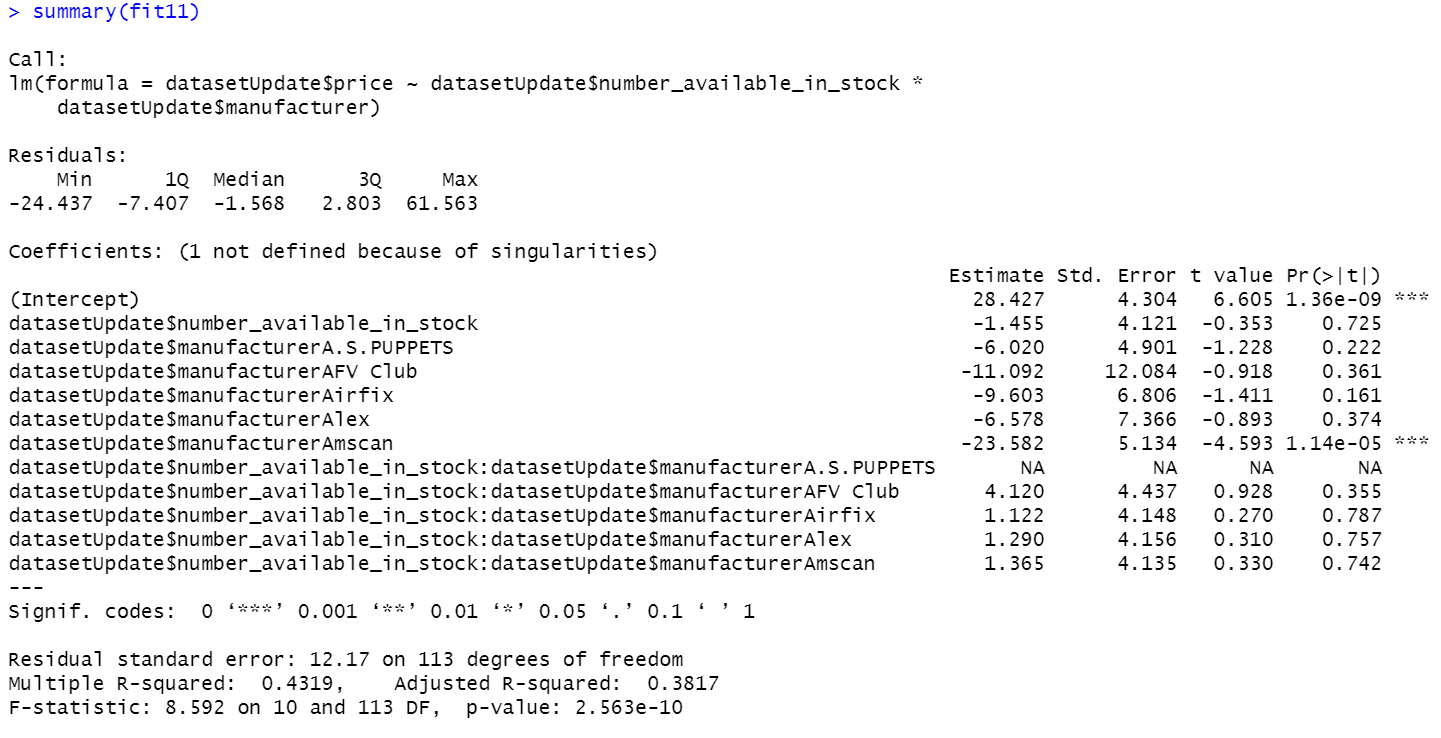
קבוצה הבסיס עבור main sale country היא: Israel

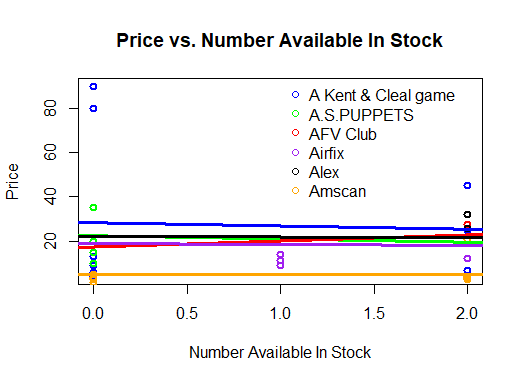
* 1. **הגדירו והוסיפו משתני אינטראקציה:**

**בדיקת משתנה אינטראקציה בין number available in stock לבין manufacturer:**

החלטנו לבדוק האם יש להכניס משתנה אינטראקציה זה מפני שצפינו כי ל manufacturer תהיה השפעה על השיפוע של number available in stock.  
ניתן לראות כי עבור 6 האופציות:

A Kent & Cleal game, A.S.PUPPETS, AFV Club, Airfix, Alex, Amscan  
רמת המובהקות של משתנה האינטראקציה אניה מובהקת ולכן **לא נכניס** את משתנה האינטראקציה למודל שלנו.  
(הערה: קיבלנו ערך אחד ויחיד שקטן מ – 0.05 עבור **החותך** של קטגורית ה Amscan)





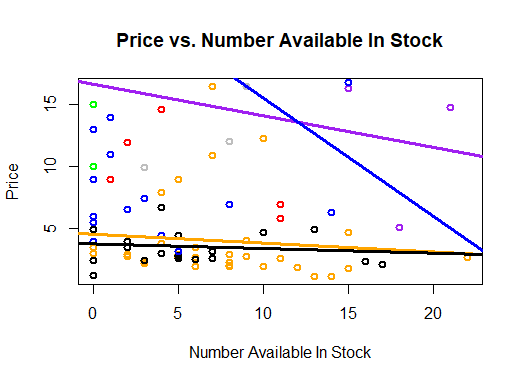
**בדיקת משתנה אינטראקציה בין number available in stock לבין eBay category:**

רצינו לבחון האם הקשר שבין number available in stock לבין ה eBay category משפר את מודל הרגרסיה מפני שאנחנו מנחים וחושבים שיש קשר בין שני המשתנים.   
ניתן לראות כי עבור eBay category יש 7 אופציות והם:

Toy, Hand Puppets, Tanks, Military, Children’s Craft Kits, Streamers, Balloons  
רמת המובהקות של משתנה האינטראקציה אניה מובהקת ולכן **לא נכניס** את משתנה האינטראקציה למודל שלנו.

תמונה שמכילה טקסט

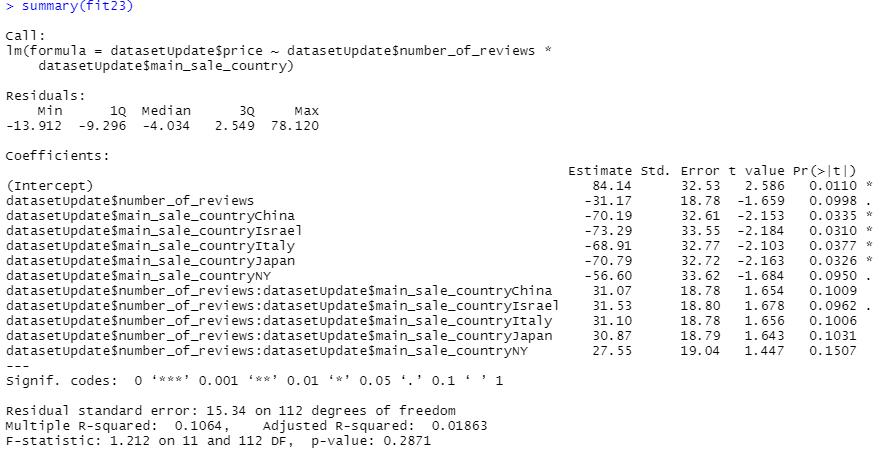
התיאור נוצר באופן אוטומטי

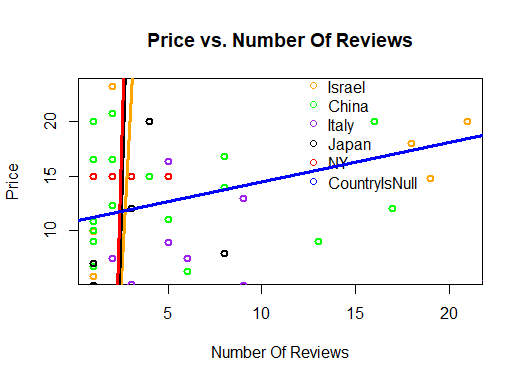


**בדיקת משתנה אינטראקציה בין number of reviews לבין main sale country:**

רצינו לבדוק את הקשר שבין number of reviews לבין ה main sale country האם הוא כן משפר את מודל הרגרסיה או שהוא לא משפר המודל, מכיוון שאנחנו מנחים וחושבים שיש קשר בין שני המשתנים.   
ניתן לראות כי עבור main sale country יש 6 אופציות לפי טבלת הנתונים שקיבלנו והם:

Null, Israel, Italy, Japan, China, NY (כלומר אין ערך בשדה בטבלת הנתונים)  
רמת המובהקות של משתנה האינטראקציה אניה מובהקת ולכן **לא נכניס** את משתנה האינטראקציה למודל שלנו.





בסוף החלטנו לעשות כל האפשריות האפשרות עבור כל המשתנים הרציפים והמשתנים הקטגוריאליים (יש לנו 3 משתנים רציפים ו 3 משתנים קטגוריאליים כלומר יש לנו בסוף 9 מודלים שונים), אחרי שבצענו בדיקה מקיפה לכלל המשתנים ואינטראקציות האפשריות במערך (כל האפשריות כתובות בקובץ ה R) ולפי תוצאות הניתוח ב R, הגענו למסקנה שאף מודל אינטראקציה לא יצא מובהק כלומר ה **Pvalue** גדולה מ 0.05, זה היה צפוי כיוון שהמתאמים שמצאנו בסעיפים קודמים לא הראו קורלציה סבירה בין המשתנים המסבירים לבין המשתנה המוסבר.

**3. התאמת המודל ובדיקת הנחות המודל:**

**3.1 בחירת משתני המודל:**

השתמשנו בשלושת אלגוריתמים לבחירת מודל הרגרסיה והם: רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים.

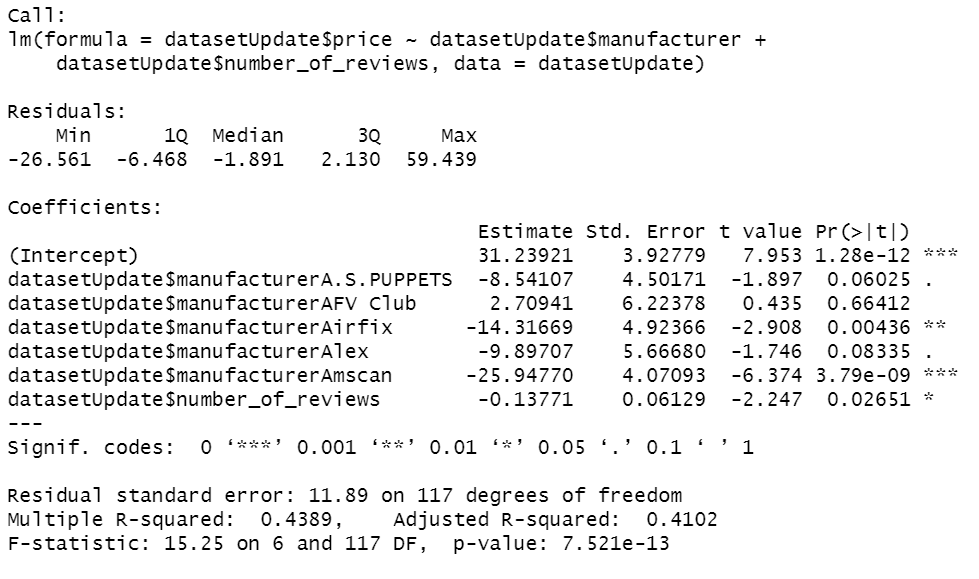
הקריטריון (המדד) בהו השתמשנו באלגוריתמים הוא מדד ה – **AIC** וממד ה- **BIC.**

מדד ה – AIC/BIC: בוחן את טיב ההתאמה של המודל לנתונים מבחינת הנראות שלו, כלומר ככל שהנראות גדולה יותר כך ערך המדד נמוך יותר. בנוסף, המדד קונס את המודל על פי מספר הפרמטרים שבו.

בהינתן מספר מודלים אופציונאליים, יבחר המודל בעל ערך מדד ה – AIC המינימאלי.

עבור מדד ה- AIC קבלנו אותו ערך בדיוק ב 3 האלגוריתמים וגם קיבלנו אותו מודל:

|  |  |  |
| --- | --- | --- |
|  | **ערך מדד ה- AIC** | **ערך מדד ה- BIC** |
| **רגרסיה לפנים** | 974.6722 | 997.2344 |
| **רגרסיה לאחור** | 974.6722 | 997.2344 |
| **רגרסיה בצעדים** | 974.6722 | 997.2344 |

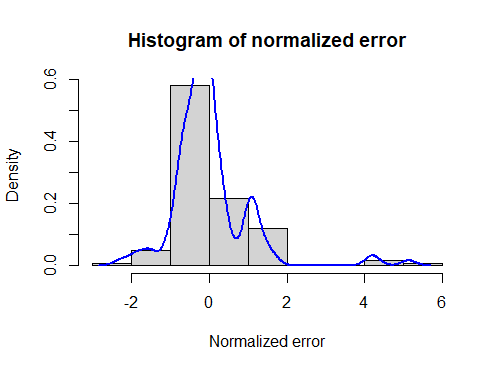
מכיוון שערך מדד ה- AIC יותר קטן מערך מדד ה- BIC אז אנחנו נמשיך עם מדד ה- AIC, המודל שקיבלנו ב **3** האלגוריתמים:

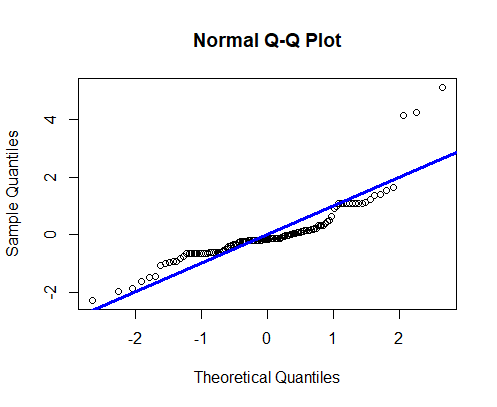
**משתנים במודל:**

|  |  |
| --- | --- |
| **בפנים** | **בחוץ** |
| **Number of reviews** | **Number available in stock** |
| **eBay category** |
| **Manufacturer** | **Average review** |
| **Main sale country** |

**3.2 בדיקת הנחות המודל:**

**1. הנחת הנורמאליות של השגיאות:**

על מנת לבדוק את ההנחה זאת נתחיל בבחינת תרשים כמותונים נורמאלי **QQ plot** ותרשים ההיסטוגרמה:



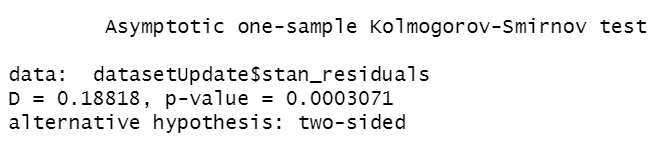
לפני תרשים ה- QQ plot ניתן לראות כי הנתונים אינם מגיעים מהתפלגות הנורמלית מפני שהם לא מסודרים בצורה טובה על הקו הישר.

ההיסטורגמה לא נותנת לא ידע מספיק והיא לא מחזיקה את הנתונים של תרשים QQ plot.

על מנת להיות בטוחים שאכן לא מתקיימת הנחת הנורמאליות של השגיאות נבצע שני מבחנים סטטיסטיים.

**מבחן KS:**

מבחן זה בודק את השערת האפס (הנתונים מגיעים מהתפלגות נורמלית -השגיאות המתוקננות מתפלגות נורמלית- אל מול ההשערה האלטרנטיבית -ההתפלגות אינה נורמלית-).

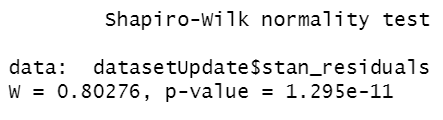
לכן על מנת לוודא את הנחת הנורמליות נרצה **Pvalue** גדול מ 0.05.

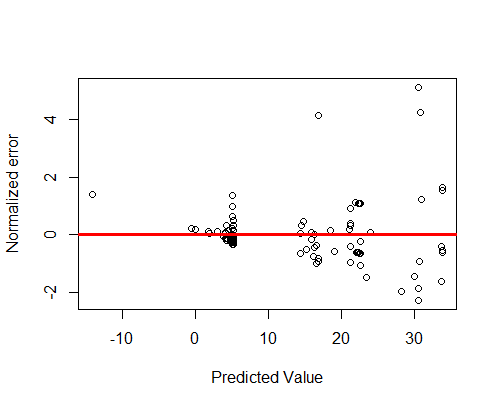
ניתן לראות מפלט ה- R שערך ה- **Pvalue** קטן מ 0.05 ולכן **כן** נדחה את השערת האפס ונאמר שברמת מבוהקות של 5% הנחת הנורמליות של השגיאות **לא** מתקיימת.

**מבחן Shapiro Wilks:**

כמו **מבחן KS**, גם המבחן בודק את השערת האפס.

גם במבחן Shapiro Wilks ערך ה- **Pvalue** קטן מ 0.05 ולכן שוב נגיד שהנחת הנורמליות **לא** מתקיימת.



**2. בדיקת הנחת שיווין שונויות:**

הנחת שוויון השונויות מתקיימת כאשר ישנו פיזור אחיד של השאריות סביב הערך 0.

מהתבוננות בגרף, ניתן לראות כי השאריות מפוזרות סביב הערך 0 אך לא באופן אחיד, בצד השמאלי נראה שהפיזור יותר אחיד סביב הערך 0 מהצד הימני. ניתן לראות שהשגיאות בתרשים הפיזור נעות בין -2 עד 2 בערך (רק 4 שגיאות מחוץ לתחום זה), ולכן, נראה כי הנחת שוויון שונויות **כן** מתקיימת.

**3. בדיקת הנחת הלינאריות:**

הנחת הלינאריות מתקיימת כאשר ניתן לראות כי השאריות מפוזרות באופן היוצר מגמתיות.

לא ניתן להצביע על מגמתיות מסוימת מהתבוננות בגרף. לכן, נראה כי ניתן להניח שהנחת הלינאריות **כן** מתקיימת.

**4. שיפור המודל:**

בסעיף זה צריכים לבדוק אם ניתן לשפר המודל על ידי שינוי בטרנספורמציה, הוספת משתנה טרנספורמציה או החלפת משתנה שכבר קיים בטרנספורמציה של אותו המשתנה.

נשתמש במדד **AIC** שכל עד ערכו יותר קטן, המודל יותר טוב ויותר מתאים לנתונים.

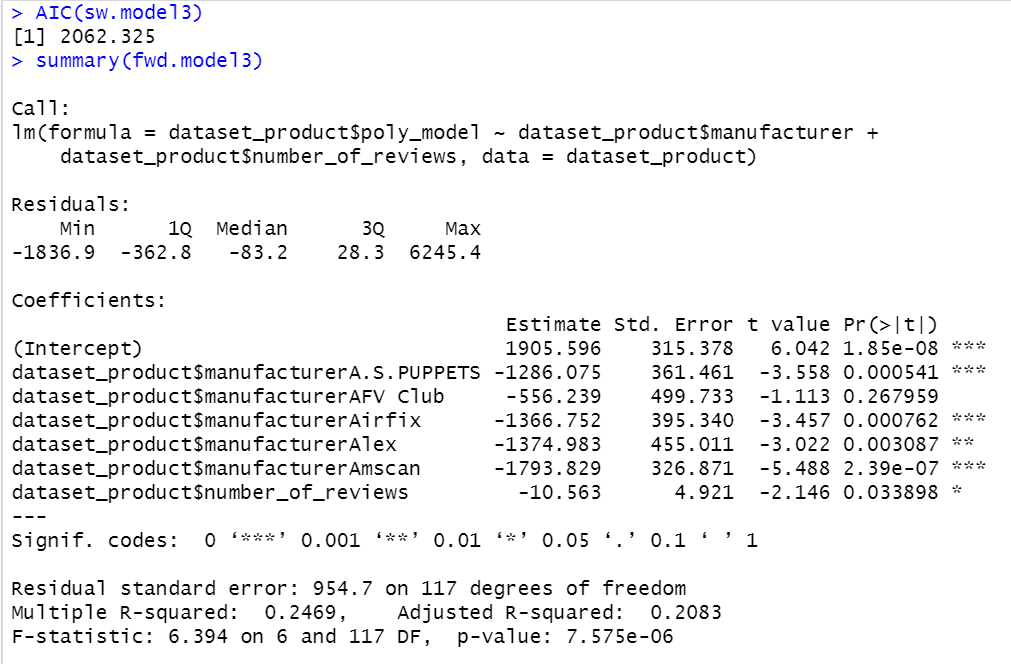
וגם נשתמש במדד טיב התאמה **Radj2** ככל שערכו יותר גדול, המודל יותר טוב ומתאים לנתונים.

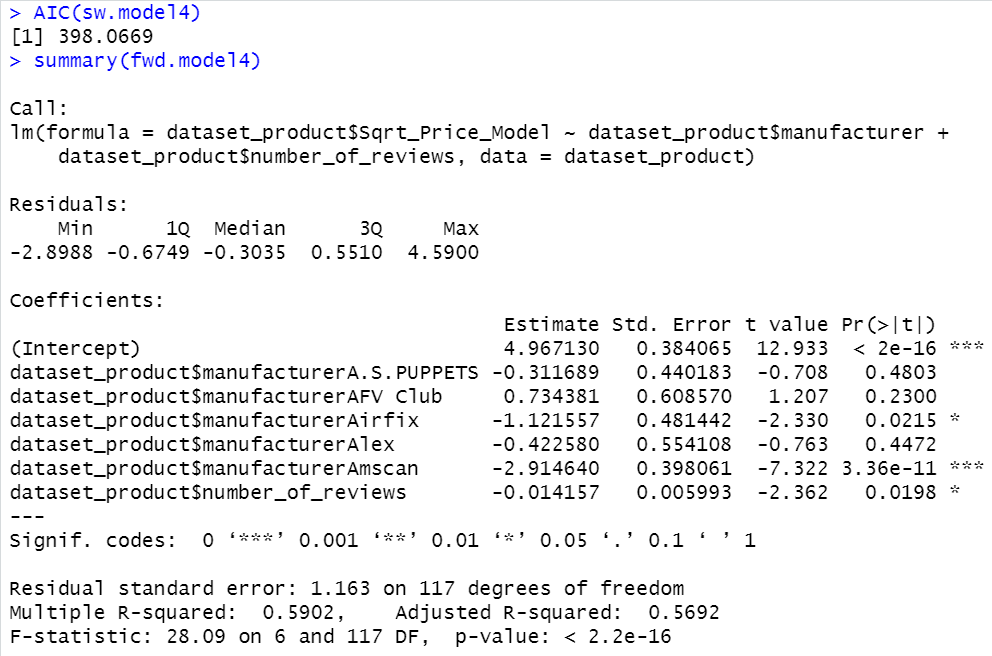
הערך של מדד ה- **AIC = 974.6722** (לפי סעיף 3.1) והערך של ה- **R2adj = 0.4102**.

גם בסעיף זה אנחנו השתמשנו ב 3 האלגוריתמים: רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים כדי לקבל המודל הכי טוב.

בגלל **שהנחת הנורמאליות של השגיאות** לא מתקיימת אז עשינו טרנספורמציה עבור המשתנה המוסבר בלבד, כלומר המחיר, אבל בבדיקה בראשונה עשינו דבר שונה, עשינו בדיקה עבור כל המשתנים המסבירים שהיו במודל מהתחלה כלומר, (החזרנו number of answered questions למודל) וקיבלנו בדיוק אותו מודל שקיבלנו בסעיף 3.1 עם אותו ערך של מדד ה- **AIC** ו **R2adj**.

אחרי כך עשינו עוד בדיקה אבל עכשיו עשינו טרנספורמציה על המשתנה המוסבר (להעלות המשתנה המוסבר בריבוע - **price ^ 2)** וקיבלנו אותו מודל בדיוק כמו בסעיף 3.1 בשלושת האלגוריתמים אבל הערך של מדד ה- AIC גדל והוא עכשיו שווה ל- **AIC = 2062.325** וערך ה- **R2adj = 0.2083** כלומר המודל המוקי יותר טוב ממודל זה (ערך מדד ה- **AIC = 974.6722** יותר קטן וערך ה- **R2adj = 0.4102** יותר גדול).



אחרי כך עשינו עוד בדיקה (בדיקה מספר 2 על המשתנה המוסבר, ועכשיו עשינו את השורש של המחיר, כלומר **price ^ 0.5**) וקיבלנו אותו מודל בדיוק כמו בסעיף 3.1 בשלושת האלגוריתמים אבל הערך של מדד ה- AIC קטן (יותר טוב) והוא עכשיו שווה ל- **AIC = 398.0669** וערך ה- **R2adj = 0.5692** כלומר המודל החדש יותר טוב ממודל המקורי (ערך מדד ה- **AIC = 974.6722** יותר גדול וערך ה- **R2adj = 0.4102**  יותר קטן).

בשלב האחרון עשינו עוד בדיקה (בדיקה מספר 3 על המשתנה המוסבר, ועכשיו הפעלנו את פונקציית ה Ln על המשתנה הסובר, כלומר **Ln(price)**) וקיבלנו מודל שונה מהמודל שקיבלנו בסעיף 3.1 בשלושת האלגוריתמים קיבלנו אותו מודל (שונה ממודל המקורי) אבל הערך של מדד ה- AIC קטן (יותר טוב) והוא עכשיו שווה ל **AIC = 231.9253** והערך של ה- **R2adj = 0.6867.**

תמונה שמכילה טקסט

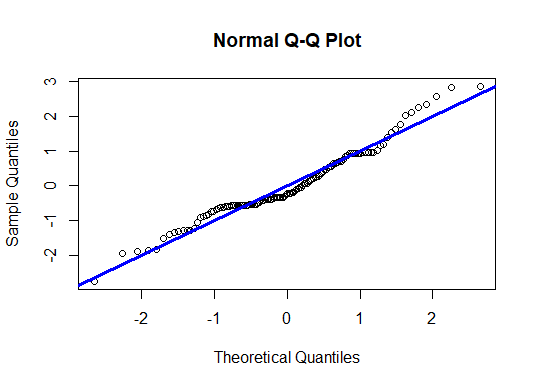
התיאור נוצר באופן אוטומטיכלומר המודל החדש יותר טוב ממודל המקורי (ערך מדד ה- **AIC = 974.6722** יותר גדול וערך ה- **R2adj = 0.4102** יותר קטן) והוא יותר טוב מהמודל בבדיקה השנייה **price ^ 0.5)).**

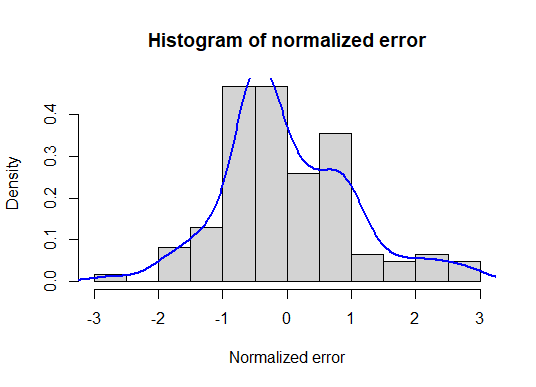
תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

לכן אנחנו נמשיך עם המודל האחרון שקיבלנו, נבדוק האם המודל הקיים את הנחות המודל:

**הנחת הנורמאליות של השגיאות:**

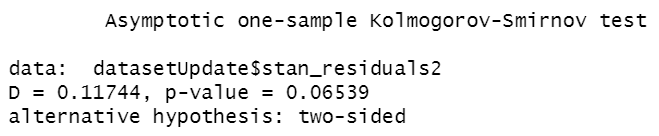




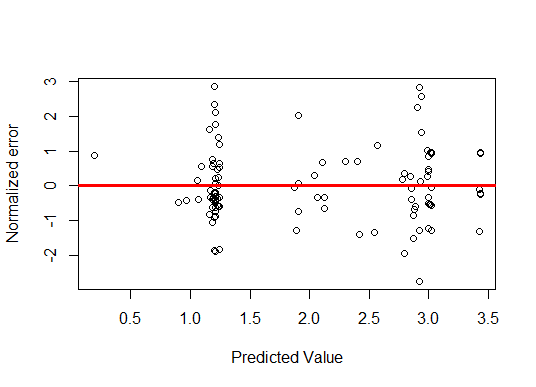
תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

לפי תרשימי הפיזור ולפי מבחן **Shapiro Wilks** הנחת הנורמאליות של השגיאות **לא** מתקיימת.



לפי תרשימי הפיזור ולפי מבחן **KS** הנחת הנורמאליות של השגיאות **כן** מתקיימת.



**בדיקת הנחת שיווין שונויות והנחת הלינאריות:**

הנחת שוויון השונויות מתקיימת כאשר ישנו פיזור אחיד של השאריות סביב הערך 0.

מהתבוננות בגרף, ניתן לראות כי השאריות מפוזרות סביב הערך 0 אך לא באופן אחיד, בצד השמאלי נראה שהפיזור יותר אחיד סביב הערך 0 מהצד הימני. ניתן לראות שהשגיאות בתרשים הפיזור נעות בין 3- עד 3 בערך ולכן, נראה כי הנחת שוויון שונויות **כן** מתקיימת.

הנחת הלינאריות מתקיימת כאשר ניתן לראות כי השאריות מפוזרות באופן היוצר מגמתיות.

לא ניתן להצביע על מגמתיות מסוימת מהתבוננות בגרף. לכן, נראה כי ניתן להניח שהנחת הלינאריות **כן** מתקיימת.

אבל לפני ערך מדד ה- AIC כן אנחנו מעדיפים המודל האחרון על המודל המקורי.