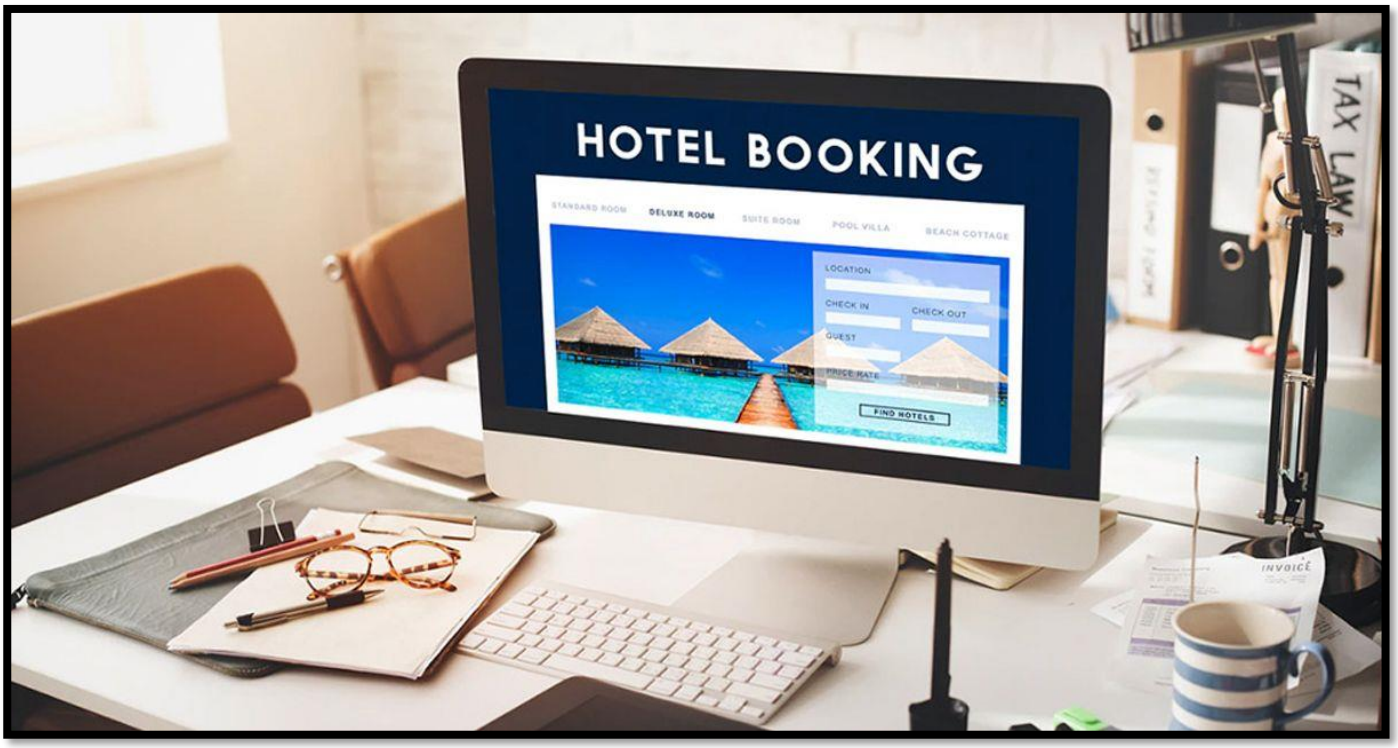
פרויקט בקורס רגרסיה ליניארית

חלק ב'

קבוצה 10

שני גואטה – 208499186

רועי עזראי – 206118754



תוכן עניינים

2. עיבוד מקדים: 3

2.1 הסרה של משתנים: 3

2.2 התאמת משתנים: 7

2.3 הגדרת משתנה דמה: 7

2.4 משתני אינטראקציה:..... 8

3. התאמת המודל ובדיקת הנחות המודל:.....10

3.1 בחירת משתני המודל: 10

3.2 בדיקת הנחות המודל: 11

4. שיפור המודל: 13

נספחים.....15

אינטראקציות נוספות שנבדקו : 15

בניית המודל :.....17

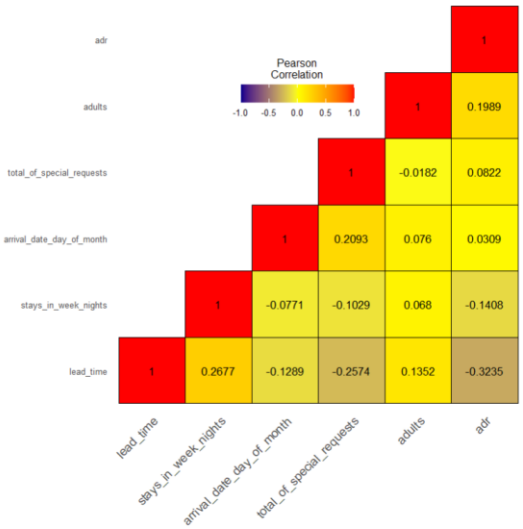
בדיקת הנחות מודלים לשיפור : 23

2. עיבוד מקדים:

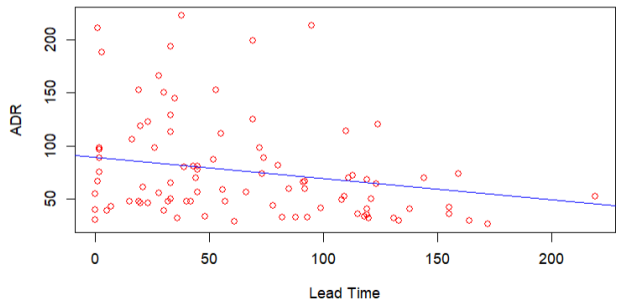
2.1 הסרה של משתנים:

על מנת לבחון האם יש להסיר חלק מהמשתנים במודל נשתמש במקדם המתאם של פירסון ובמובהקות תוצאה-P Value עבור משתנים רציפים, ועבור משתנים קטגוריאלים נשתמש במובהקות תוצאהP-Value בתרשימי Boxplot.
בנוסף נרצה לבחון הסרה של משתנה מסביר אם קיים קשר לינארי חזק בינו לבין משתנה מסביר נוסף כיוון שנרצה עבור כל משתנה במודל ערך שונה וייחודי לו.
מקדם המתאם של פירסון עוזר לבחון את הקשר הלינארי בין 2 משתנים רציפים. כאשר ערכו מתקרב ל-1 בערך מוחלט ניתן לומר שיש קשר לינארי חזק בין המשתנים, כאשר ערכו מתקרב לאפס ניתן לומר שהקשר חלש בין המשתנים.
מובהקות התוצאה P-Value הינה ההסתברות לקבל תוצאה זהה או קיצונית לפחות כמו זו שהתקבלה בניסוי תחת השערת האפס (אשר מחזקת את ההשערה האלטרנטיבית). במודל הרגרסיה המרובה נבדוק באיזה מידה המשתנה שנבחן מסביר את המוסבר.

נתבונן בגרף הבא המייצג את ערך מקדם המתאם בין המשתנים המסבירים למשתנים המסבירים ולמשתנה המוסבר (adr) :

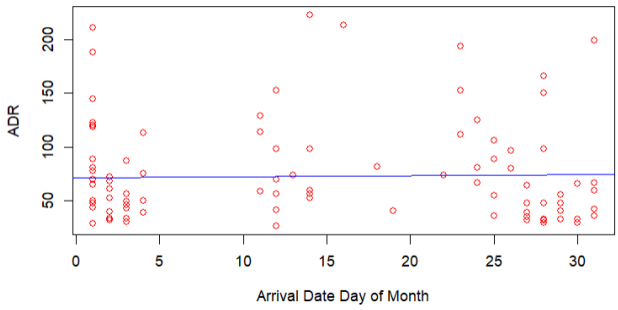


המשתנה המוסבר (**adr**) אל מול (**X1 - Lead Time**) :
בין המשתנה המוסבר למשתנה זה יש קורלציה -0.3235 המעידה על קשר לינארי שלילי יחסית חלש על פי מדד פירסון אך ביחס לשאר המשתנים במודל שלנו זהו המתאם החזק ביותר ביחס למשתנה המוסבר. מתוך הגרף קשה לראות האם המתאם בין המשתנים טוב.
בהסתכלות על מודל הרגרסיה התקבל P-Value=0.000233 נמוך מאוד וזה מחזק את הטענה שיתכן ויש קשר בין המשתנים. ציפינו שיהיה קשר בין מספר הימים מתאריך ההזמנה לתאריך ההגעה(1X) לתעריף היומי הממוצע(Y) כי יתכן שהביקוש לתאריכים רחוקים יותר יהיה נמוך בהשוואה לתאריכים הקרובים יותר למועד ההגעה ובהתאם לרמת הביקוש המחיר יקבע וישפיע על תעריף הממוצע היומי. לכן נבחר להשאיר את משתנה זה במודל.




```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               89.49907    5.73155  15.615  < 2e-16 ***
datasetCorNum$lead_time -0.20113    0.05304  -3.792 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.68 on 123 degrees of freedom
Multiple R-squared:  0.1047,    Adjusted R-squared:  0.0974
F-statistic: 14.38 on 1 and 123 DF,  p-value: 0.0002329
```

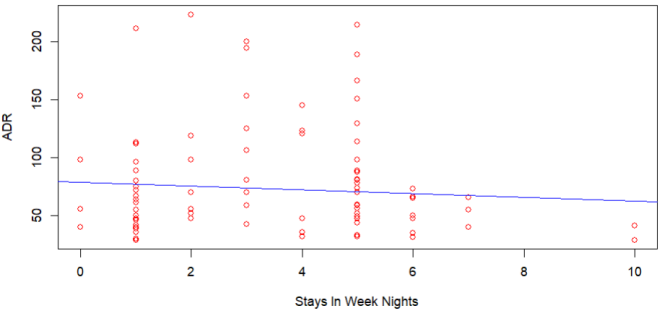
המשתנה המוסבר (**adr**) אל מול (**X3 - Arrival Date Day Of Month**):



בין המשתנה המוסבר למשתנה זה יש קורלציה 0.0309. מכך ומתוך הגרף לא ניתן לזהות מתאם בין המשתנים. בהסתכלות על מודל הרגרסיה התקבל P-Value=0.733 גבוה. נתונים אלו מחזקים את מה שציפינו לראות כיוון שיום ההגעה בחודש תלוי גם בחודש עצמו. יש חודשים שתחילת החודש יהיה יותר מבוקש ויש כאלו שבסופו והסתכלות על הימים בלבד מתעלמת מהמידע החשוב שהוא החודש עצמו ולא מסבירה נכון את המשתנה המוסבר. לכן נבחר להוריד את משתנה זה מהמודל כיוון שהנתונים מעידים על כך שאינו מסביר טוב את המשתנה המוסבר ואין קשר סיבתי ניכר.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      71.1004     5.4048  13.155   <2e-16 ***
dataset$arrival_date_day_of_month  0.1118     0.3266   0.342    0.733
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.92 on 123 degrees of freedom
Multiple R-squared:  0.0009522, Adjusted R-squared:  -0.00717
F-statistic: 0.1172 on 1 and 123 DF,  p-value: 0.7326
```

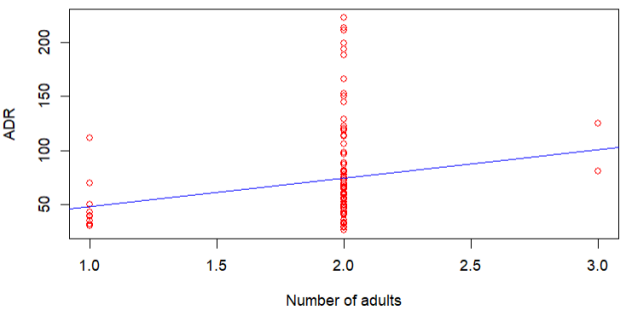
המשתנה המוסבר (**adr**) אל מול (**X4 - Stays In Week Nights**) :



בחלק א' של הפרויקט ביצענו הוצאת חריגים ממשתנה זה וכעת נבחן את ההשפעה שלו על המשתנה המוסבר בהיעדר התצפיות החריגות. כעת בין המשתנה המוסבר למשתנה זה יש קורלציה -0.1408 המעידה על קשר ליניארי שלילי חלש מאוד. מכך ומתוך הגרף לא ניתן לזהות מתאם בין המשתנים. בהסתכלות על מודל הרגרסיה התקבל P-Value=0.354 גבוה שמחזק את הטענה שמשתנה זה לא מסביר בצורה טובה את המשתנה המוסבר ולדעתנו קשה להבחין בקשר סיבתי בין מספר לילות השבוע(שני עד שישי) 4X לבין התעריף היומי הממוצע (Y). לכן נבחר להסיר את משתנה זה מהמודל.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 78.924      7.240  10.901   <2e-16 ***
bpAWNX$stays_in_week_nights -1.651      1.774  -0.931    0.354
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.92 on 120 degrees of freedom
Multiple R-squared:  0.007168,  Adjusted R-squared:  -0.001105
F-statistic: 0.8664 on 1 and 120 DF,  p-value: 0.3538
```

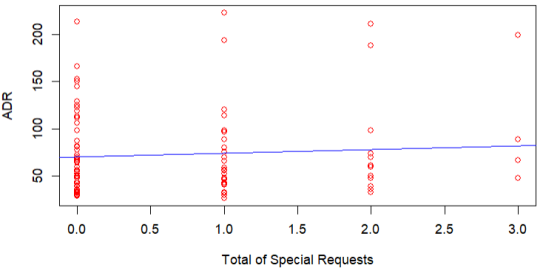
המשתנה המוסבר (**adr**) אל מול (**X5 - Adults**) :



בין המשתנה המוסבר למשתנה זה יש קורלציה 0.1989 המעידה על קשר ליניארי חלש על פי מדד פירסון אך ביחס לשאר המשתנים במודל שלנו זהו המתאם השני בחוזקתו מול המשתנה המוסבר. מהגרף ניתן לראות שככל שמספר המבוגרים גדל, כך תעריף הממוצע היומי עולה המעיד על קשר ליניארי חיובי. בהסתכלות על מודל הרגרסיה התקבל P-Value=0.0261 נמוך וזה מחזק את הטענה שיתכן ויש קשר בין המשתנים. ציפינו שיהיה קשר בין מספר המבוגרים בחדר(5X) לתעריף היומי הממוצע(Y) כיוון שככל שיש יותר מבוגרים בחדר המחיר עולה וכך גם משפיע על התעריף היומי הממוצע. לכן נבחר להשאיר את המשתנה זה במודל.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    21.67      22.84   0.949   0.3446
dataset$adults 26.33      11.69   2.251   0.0261 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.1 on 123 degrees of freedom
Multiple R-squared:  0.03958,   Adjusted R-squared:  0.03177
F-statistic: 5.069 on 1 and 123 DF,  p-value: 0.02614
```

המשתנה המוסבר (**adr**) אל מול (**X9 - Total Of Special Requests**) :



בין המשתנה המוסבר למשתנה זה יש קורלציה 0.0822 המעידה על קשר לינארי חלש מאוד. מתוך הגרף לא ניתן לזהות מתאם בין המשתנים. בהסתכלות על מודל הרגרסיה התקבל P-Value=0.362, גבוה שמחזק את הטענה שמשתנה זה לא מסביר בצורה טובה את המשתנה המוסבר ולא ניתן לראות קשר סיבתי בין מספר הבקשות המיוחדות של הלקוח לתעריף הממוצע היומי. לכן נבחר להסיר את משתנה זה מהממודל .

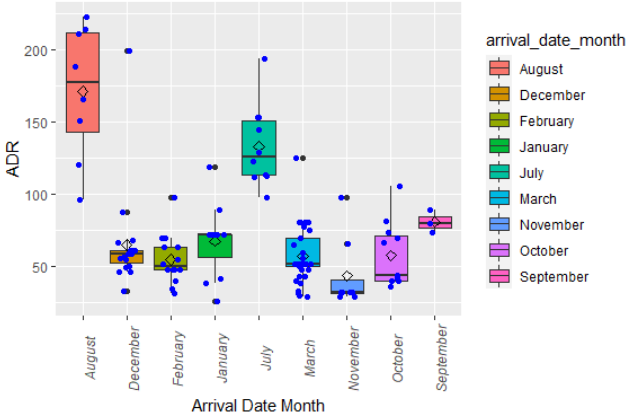
```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                       69.988      4.596  15.228   <2e-16 ***
dataset$total_of_special_requests  4.075      4.456   0.915    0.362
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.8 on 123 degrees of freedom
Multiple R-squared:  0.006756,  Adjusted R-squared:  -0.00132
F-statistic: 0.8366 on 1 and 123 DF,  p-value: 0.3622
```

<u>משתנים קטגוריאלים:</u>

המשתנה המוסבר (**adr**) אל מול (**X2 - Arrival Date Month**) :

ניתן לראות מתרשים ה Boxplot שוני במדד הממוצע והחציון בין החודשים השונים ביחס למשתנה המוסבר. בנוסף ערך ה P-Value נמוך מאוד ומעיד על כך שהמשתנה מסביר בצורה טובה את המשתנה המוסבר. הנתונים האלו מסתדרים עם ההבנה שבחודשים שונים יש ביקושים ותמחור שונה של החופשה במלון ומכך שחודש החופשה משפיע על התעריף היומי הממוצע. לכן נבחר שלא להסיר את משתנה זה מהמודל.

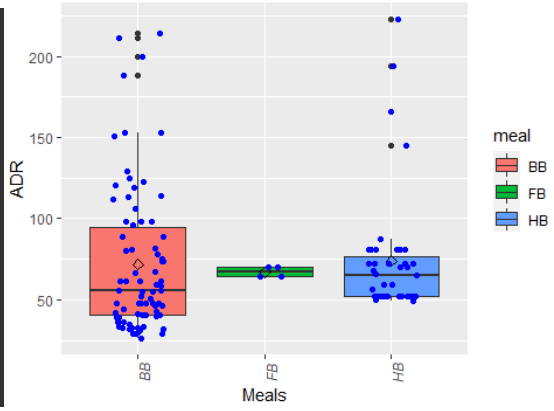
```
Call:
lm(formula = dataset$adr ~ dataset$arrival_date_month, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-74.955 -14.612  -5.612  11.581 134.595

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                            171.255      9.142  18.732  < 2e-16 ***
dataset$arrival_date_monthDecember    -106.120     10.898  -9.737  < 2e-16 ***
dataset$arrival_date_monthFebruary    -116.070     11.460 -10.128  < 2e-16 ***
dataset$arrival_date_monthJanuary     -103.408     12.015  -8.606 4.26e-14 ***
dataset$arrival_date_monthJuly         -37.836     12.266  -3.085  0.00255 **
dataset$arrival_date_monthMarch       -113.643      9.994 -11.371  < 2e-16 ***
dataset$arrival_date_monthNovember    -127.303     12.929  -9.846  < 2e-16 ***
dataset$arrival_date_monthOctober     -113.384     12.015  -9.437 4.96e-16 ***
dataset$arrival_date_monthSeptember    -90.208     17.506  -5.153 1.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.86 on 116 degrees of freedom
Multiple R-squared:  0.6415,    Adjusted R-squared:  0.6168
F-statistic: 25.94 on 8 and 116 DF,  p-value: < 2.2e-16
```



המשתנה המוסבר (**adr**) אל מול (**X6 - Meal**) :

ניתן לראות מתרשים ה Boxplot שאין הבדלים משמעותיים בערך החציון והממוצע בין סוגי הפנסיון השונים אך ערך ה P-Value גבוה מאוד ומעיד על כך שהמשתנה מסביר בצורה טובה את המשתנה המוסבר. למרות שתרשים הפיזור לא מראה קשר מובהק בין המשתנה הזה למשתנה המוסבר, אנחנו חושבים שלסוג הפנסיון יש השפעה על התעריף היומי הממוצע כי כל סוג פנסיון מתומחר שונה, עם זאת נרצה לאחד בין סוג פנסיון FB ו HB כיוון שעבורם הממוצע והחציון דומים במיוחד ותרומתם למשתנה המוסבר דומה. לכן נבחר שלא להסיר את משתנה זה מהמודל.

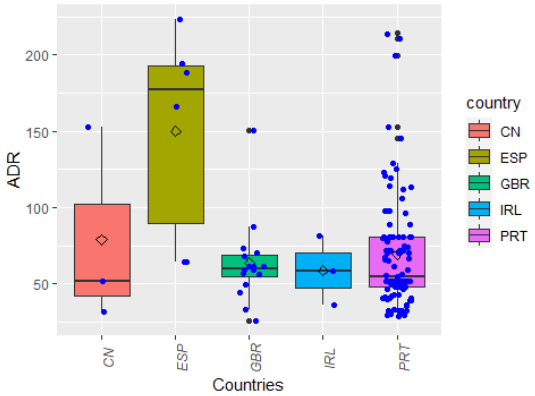
```
Call:
lm(formula = dataset$adr ~ dataset$meal, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-45.962 -23.882 -13.182   7.061 149.061

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      71.882      4.765  15.084   <2e-16 ***
dataset$mealFB   -4.882     21.576  -0.226    0.821
dataset$mealHB    2.057      7.994   0.257    0.797
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.09 on 122 degrees of freedom
Multiple R-squared:  0.001106,  Adjusted R-squared:  -0.01527
F-statistic: 0.06753 on 2 and 122 DF,  p-value: 0.9347
```



המשתנה המוסבר (**adr**) אל מול (**X7 - Country**) :

ניתן לראות מתרשים ה Boxplot שאין הבדלים משמעותיים בערך החציון והממוצע בין המדינות השונות מלבד ספרד(ESP) אך ערך ה P-Value נמוך מאוד ומעיד על כך שהמשתנה מסביר בצורה טובה את המשתנה המוסבר. למרות שתרשים הפיזור לא מראה קשר מובהק בין המשתנה הזה למשתנה המוסבר, אנחנו חושבים שלמדינה יש השפעה על התעריף היומי הממוצע כי לכל מדינה יש כלכלה שונה ותמחור שונה של החופשה, עם זאת נרצה לאחד בין סוגי המדינות אירלנד (IRL) ובריטניה (GBR) כיוון שלהן אופי ואורך החיים דומה ותרומתם למשתנה המוסבר תהיה דומה.לכן נבחר שלא להסיר את משתנה זה מהמודל.

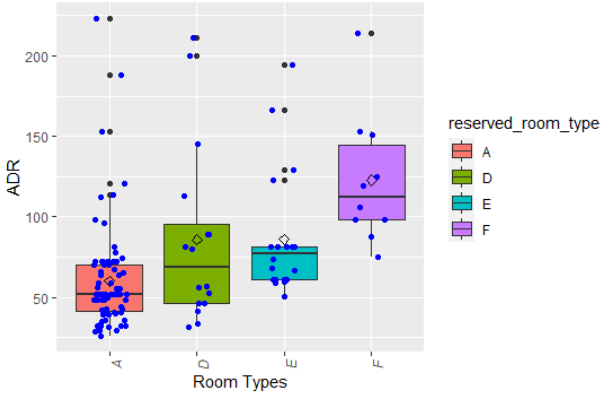
```
Call:
lm(formula = dataset$adr ~ dataset$country, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-85.905 -21.332  -7.613  11.668 144.668 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)           78.817     22.202   3.550 0.000551 ***
dataset$countryESP    71.088     27.191   2.614 0.010085 *  
dataset$countryGBR   -15.254     24.194  -0.630 0.529581    
dataset$countryIRL   -20.137     31.398  -0.641 0.522525    
dataset$countryPRT    -9.484     22.542  -0.421 0.674699    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.45 on 120 degrees of freedom
Multiple R-squared:  0.1798,	Adjusted R-squared:  0.1524 
F-statistic: 6.575 on 4 and 120 DF,  p-value: 8.079e-05
```



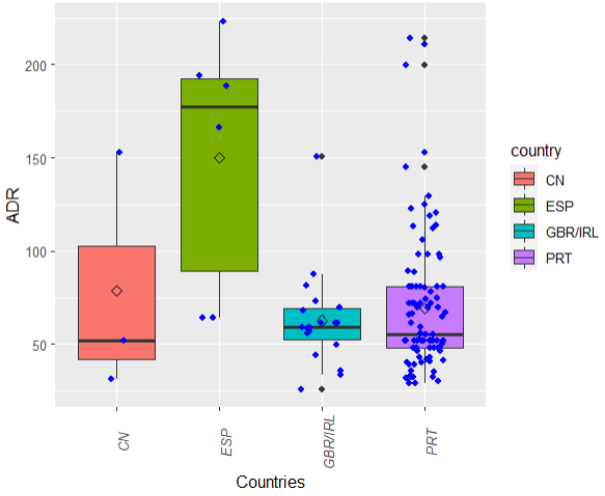
המשתנה המוסבר (**adr**) אל מול (**X8 - Reserved Room Type**) :

ניתן לראות מתרשים ה Boxplot שוני במדד הממוצע והחציון בין סוגי החדרים השונים ביחס למשתנה המוסבר. בנוסף ערך ה P-Value נמוך מאוד ומעיד על כך שהמשתנה מסביר בצורה טובה את המשתנה המוסבר. הנתונים האלו מסתדרים עם ההבנה שעבור חדרים שונים התמחור שונה ומשפיע על התעריף היומי הממוצע ולכן נבחר שלא להסיר את משתנה זה מהמודל.

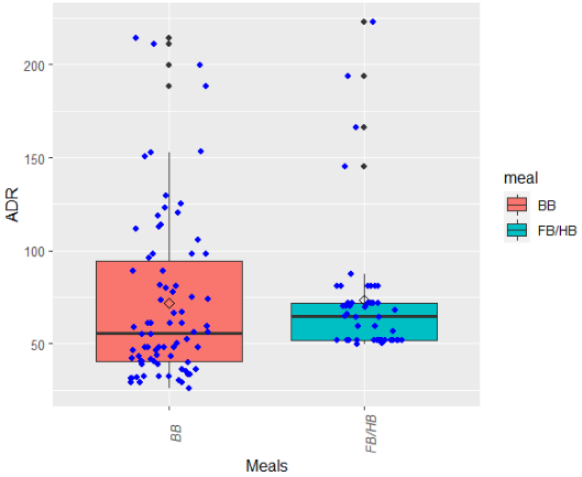
```
Call:
lm(formula = dataset$adr ~ dataset$reserved_room_type, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-54.279 -23.954  -7.954   9.976 163.046 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     59.954      4.246  14.120  < 2e-16 ***
dataset$reserved_room_typeD     25.776     10.346   2.491  0.01408 *  
dataset$reserved_room_typeE     26.042      9.447   2.757  0.00674 ** 
dataset$reserved_room_typeF     62.670     12.667   4.947 2.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.74 on 121 degrees of freedom
Multiple R-squared:  0.2034,	Adjusted R-squared:  0.1836 
F-statistic:  10.3 on 3 and 121 DF,  p-value: 4.352e-06
```



2.2 התאמת משתנים:



עבור המשתנה 7X המייצג את המדינות בהן מתבצעת החופשה, ראינו לנכון לאחד את המדינה אירלנד (IRL) עם המדינה בריטניה (GBR). ניתן לראות בתרשים ה Boxplot בסעיף 2.1 שהממוצע והחציון בשתי מדינות דומה מאוד, בנוסף לאירלנד יש 3 תצפיות בסהכ בעלות השפעה זניחה על המודל. מכאן ומתוך ההבנה שבריטניה ואירלנד הן מדינות דומות באופיין התיירותי עקב מיקומן הגיאוגרפי ואורך החיים הדומה בשתי המדינות אנחנו חושבים שלשתיהן יש השפעה דומה על המשתנה המוסבר ונרצה לאחד אותן לקטגוריה אחת. בתרשים ה Boxplot ראינו שהממוצע והחציון של מדינת פורטוגל דומה מאוד לבריטניה ולאירלנד אך בעקבות השוני במיקום הגיאוגרפי ובתרבות בין המדינות, איחוד זה אינו נכון מבחינה רעיונית ונבחר שלא לחבר בין הקטגוריות.
מצורף תרשים Boxplot לאחר איחוד הקטגוריות:



עבור המשתנה 6X המייצג את סוג הפנסיון, ראינו לנכון לאחד את סוג הפנסיון FB עם סוג פנסיון HB. ניתן לראות בתרשים ה Boxplot בסעיף 2.1 שהממוצע והחציון עבור שתי הקטגוריות דומה מאוד. בנוסף, ישנם 4 תצפיות בלבד לסוג פנסיון FB אשר בעלות השפעה זניחה על המודל. מכאן ומתוך ההבנה שההבדל בין בחירת FB ל HB לא משמעותית לעומת בחירת פנסיון BB שבו יש רק ארוחת בוקר שהיא הארוחה הכי קלה מבין כל הארוחות וסביר להניח שהיא חבילת הבסיס בהזמנת חופשה (ללא עלות נוספת לעומת FB ו HB). אנחנו חושבים שלקטוגריות FB ו HB יש השפעה דומה על המשתנה המוסבר ונרצה לאחד אותן לקטגוריה אחת.
מצורף תרשים Boxplot לאחר איחוד הקטגוריות:

2.3 הגדרת משתנה דמה:

נגדיר משתני דמה עבור המשתנים הקטגוריאלים במודל:

עבור משתנה **X6 - Meal**:
קבוצת הבסיס תהיה BB ונגדיר משתנה דמה אחד ל FB/HB:

$$\text{FBHB} = \begin{cases} 1\,, & \text{if the meal is FB/HB} \\ 0\,, & \text{else} \end{cases}$$

עבור משתנה **X2 - Arrival Date Month**:
קבוצת הבסיס תהיה August וניצור 8 משתני דמה:

$$\text{February} = \begin{cases} 1\,, & \text{if the month is February} \\ 0\,, & \text{else} \end{cases} \qquad \text{December} = \begin{cases} 1\,, & \text{if the month is December} \\ 0\,, & \text{else} \end{cases}$$

$$\text{July} = \begin{cases} 1\,, & \text{if the month is July} \\ 0\,, & \text{else} \end{cases} \qquad \text{January} = \begin{cases} 1, & \text{if the month is January} \\ 0\,, & \text{else} \end{cases}$$

$$\text{November} = \begin{cases} 1\,, \text{if the month is November} \\ 0\,, \quad \text{else} \end{cases} \quad \text{March} = \begin{cases} 1\,, \text{if the month is March} \\ 0\,, \quad \text{else} \end{cases}$$

$$\text{September} = \begin{cases} 1\,, \text{if the month is September} \\ 0\,, \quad \text{else} \end{cases} \quad \text{October} = \begin{cases} 1, \text{if the month is October} \\ 0\,, \quad \text{else} \end{cases}$$

עבור משתנה **X7 - Country**:

קבוצת הבסיס תהיה PRT וניצור 3 משתני דמה:

$$\text{ESP} = \begin{cases} 1\,, \text{ if the country is ESP} \\ 0\,, \quad \text{else} \end{cases} \quad \text{CN} = \begin{cases} 1\,, \text{ if the country is CN} \\ 0\,, \quad \text{else} \end{cases}$$

$$\text{GBR/IRL} = \begin{cases} 1\,, \text{ if the country is GBR/IRL} \\ 0\,, \quad \text{else} \end{cases}$$

עבור משתנה **X8 - Reserved Room Type**:

קבוצת הבסיס תהיה A וניצור 3 משתני דמה:

$$\text{D} = \begin{cases} 1\,, \text{if the room type is D} \\ 0\,, \quad \text{else} \end{cases} \quad \text{E} = \begin{cases} 1\,, \text{if the room type is E} \\ 0\,, \quad \text{else} \end{cases}$$

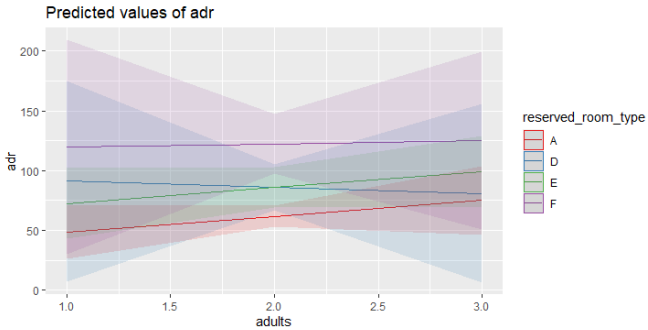
$$\text{F} = \begin{cases} 1\,, \text{if the room type is F} \\ 0\,, \quad \text{else} \end{cases}$$

2.4 משתני אינטראקציה:

<u>משתנה אינטראקציה בין X5 - Adults לבין X8 - Reserved Room Type:</u>

בחרנו לבחון משתנה אינטראקציה זה כיוון שסביר להניח שסוגי חדרים שונים יתאימו למספר שונה של מבוגרים. מהגרף ניתן לראות שההבדל בין השיפועים לא משמעותי עבור חדרים מסוג A, E ו- F. עבור חדר מסוג D יש הבדל בשיפוע ביחס לשאר סוגי החדרים אך בגלל שערך ה P-Value גבוה אנחנו נבחר שלא להוסיף את המשתנה הזה למודל שלנו.

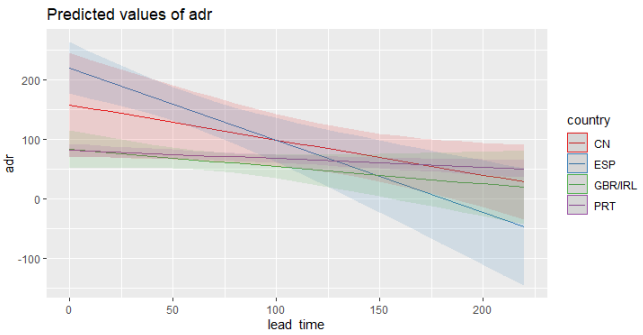
```
Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 35.034     23.389   1.498   0.1368
adults                      13.392     12.357   1.084   0.2807
reserved_room_typeD         61.100     84.841   0.720   0.4728
reserved_room_typeE         24.177      9.672   2.500   0.0138 *
reserved_room_typeF         82.046     88.183   0.930   0.3541
adults:reserved_room_typeD  -18.437     41.170  -0.448   0.6551
adults:reserved_room_typeE       NA         NA      NA       NA
adults:reserved_room_typeF  -10.752     41.943  -0.256   0.7981
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.02 on 118 degrees of freedom
Multiple R-squared:  0.2114,    Adjusted R-squared:  0.1713
F-statistic: 5.271 on 6 and 118 DF,  p-value: 0.00007614
```



<u>משתנה אינטראקציה בין X1 - Lead Time לבין X7 - Country:</u>

בחרנו לבחון משתנה אינטראקציה זה כיוון שסביר להניח שבין המדינות יש הבדל בתיירות ובביקוש וביעדים בעלי ביקוש גבוה יש צורך לבצע הזמנה של החופשה מוקדם יותר מביעדים עם ביקוש נמוך. מהגרף ניתן לראות כי קיים הבדל בין השיפועים אך משתנה האינטראקציה לא מובהק ולכן לא נוסיף את משתנה זה למודל.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               158.1446    44.3673   3.564 0.000529 ***
lead_time                  -0.5905     0.2948  -2.003 0.047503 *
countryESP                 62.4432    49.6929   1.257 0.211408
countryGBR/IRL            -74.9203    47.3118  -1.584 0.115997
countryPRT                -76.1842    44.7193  -1.704 0.091109 .
lead_time:countryESP       -0.6281     0.4200  -1.495 0.137486
lead_time:countryGBR/IRL    0.2993     0.3591   0.833 0.406277
lead_time:countryPRT        0.4468     0.2990   1.494 0.137773
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.63 on 117 degrees of freedom
Multiple R-squared:  0.3515,	Adjusted R-squared:  0.3127
F-statistic: 9.058 on 7 and 117 DF,  p-value: 0.000000006688
```



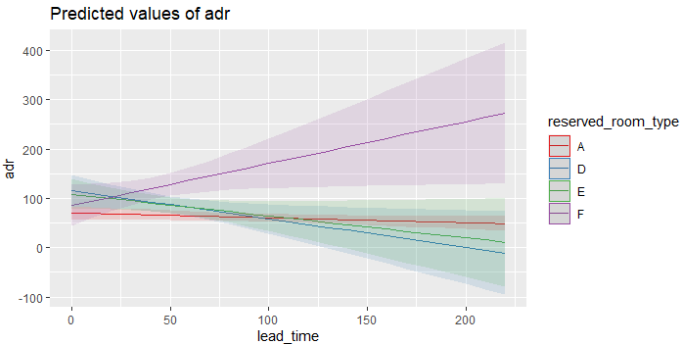
<u>משתנה אינטראקציה בין X1 - Lead Time לבין X8 - Reserved Room Type:</u>

בחרנו לבחון משתנה אינטראקציה זה כיוון שישנם חדרים בעלי ביקוש גבוה שיש צורך להזמין אותם תקופה ארוכה יותר מראש על מנת לשריין בהם מקום ולדעתנו סוג החדר משפיע על ה Lead Time.

מהגרף ניתן לראות כי בין חדרים מסוג D ו E השיפוע כמעט זהה, אך השיפוע של חדר מסוג F ו A שונה.

מהסתכלות על מודל הרגרסיה המשתנים האינטראקציה של החדרים D ו E אינם מובהקים אך רמת המובהקות של חדר D יחסית נמוכה, ולכן נבחר להכניס את משתנה האינטראקציה הזה למודל שלנו.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     70.43929    7.10912   9.908   <2e-16 ***
lead_time                       -0.09960    0.05546  -1.796   0.0751 .
reserved_room_typeD             45.31676   17.44868   2.597   0.0106 *
reserved_room_typeE             38.08922   17.19768   2.215   0.0287 *
reserved_room_typeF             15.21866   22.27025   0.683   0.4957
lead_time:reserved_room_typeD   -0.47439    0.25725  -1.844   0.0677 .
lead_time:reserved_room_typeE   -0.34353    0.26978  -1.273   0.2054
lead_time:reserved_room_typeF    0.94939    0.41205   2.304   0.0230 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.05 on 117 degrees of freedom
Multiple R-squared:  0.2971,	Adjusted R-squared:  0.255
F-statistic: 7.064 on 7 and 117 DF,  p-value: 0.0000004986
```



3. התאמת המודל ובדיקת הנחות המודל:

3.1 בחירת משתני המודל:

על מנת לבחון ולבחור את משתני המודל שלנו נשתמש באלגוריתמים של רגרסיה לאחור, רגרסיה לפנים ורגרסיה בצעדים לפי מדד AIC ולפי מדד BIC. את המודל נבחר לפי ערך הAIC המינימלי המתקבל. במצב של שוויון במדד הAIC נשתמש במדד R^2 adj כדי לבחור את המודל, נבחר ב R המקסימלי מבין כולם.

BIC Stepwise	AIC Stepwise	BIC Backward	AIC Backward	BIC Forward	AIC Forward	Full Model	
763.774	759.444	763.774	759.444	763.774	759.444	761.380	AIC
809.0272	810.354	809.0272	810.354	809.0272	810.354	820.775	BIC
0.7708	0.7816	0.7708	0.7816	0.7708	0.7816	0.7825	R^2 adj

קיבלנו את ערך ה AIC הנמוך ביותר מכל אחד מן התהליכים המנוחים לפי מדד ה AIC , למרות שבמודל המלא שלנו קיבלנו ערך R^2 adj גבוה יותר בפער יחסית מזערי (0.0009) אנחנו נבחר במודל שהתקבל ב AIC מכוון שבמודל שהתקבל הוסר רק משתנה האינטראקציה שהוספנו בסעיף הקודם, ניתן לראות שבכל אחד מן תהליכי בניית המודל , בכולם בחרו להסיר את משתנה זה ואף ב AIC Backward הוא השתנה שהוסר ראשון. מוטיבציה נוספת לקחת את המודל ללא משתנה האינטראקציה היא שבסעיף הקודם אכן ראינו שהוא אינו מובהק לחלוטין וההחלטה להוסיף אותו למודל הייתה בשל

<u>המודל המלא:</u>

```
adr ~ lead_time  + adults +
        factor(arrival_date_month)  +
        factor(country)   +
        factor(reserved_room_type)  +
        factor(meal) +
        lead_time * reserved_room_type
```

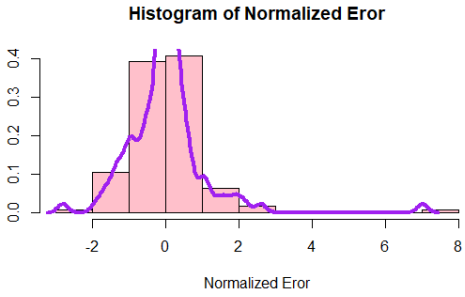
<u>המודל המתקבל:</u>

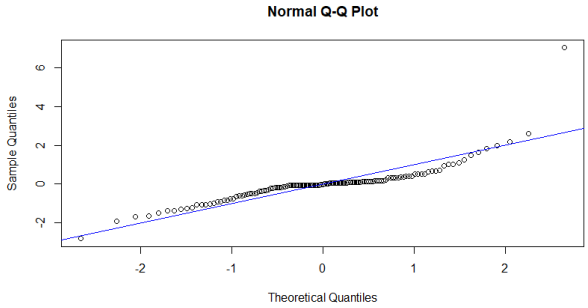
```
adr ~ lead_time  + adults +
        factor(arrival_date_month)  +
        factor(country)   +
        factor(reserved_room_type)  +
        factor(meal)
```

3.2 בדיקת הנחות המודל:

בדיקת הנחת הנורמאליות של השגיאות:

בהסתכלות על ההיסטוגרמה ועל גרף תרשים Normal Q-Q, המתאר את ההתנהגות המצופה מנתונים המגיעים מהתפלגות נורמלית, ניתן לראות שרוב התצפיות לא נמצאות על הקו הישר, כלומר שההתפלגות לא נורמלית.





מבחנים סטטיסטיים לבדיקת קיום הנחת הנורמאליות :

<u>מבחן Kolmogorov – Smirnov :</u>

בהרצת מבחן זה קיבלנו P-Value קטן מאוד ולכן נדחה את השערת האפס ברמת מובהקות 5% ונאמר שהנחת הנורמאליות לא מתקיימת במודל זה. תוצאות המבחן מחזקות את מה שראינו מהתרשימים.

```
> # KS Test - Normality :
> ks.test(x = datasetNewX$stan_residuals, y = "pnorm",
+         alternative = "two.sided", exact = NULL)

        One-sample Kolmogorov-Smirnov test

data:  datasetNewX$stan_residuals
D = 0.18257, p-value = 0.0004808
alternative hypothesis: two-sided
```

<u>מבחן Shapiro – Wilk :</u>

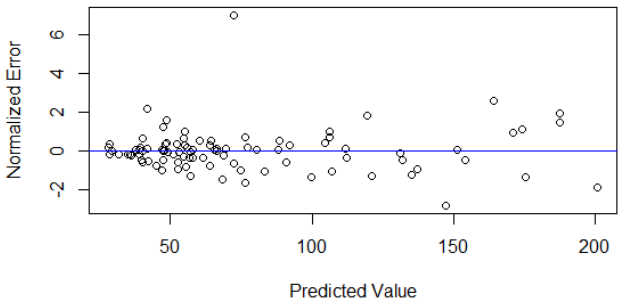
גם בהרצת מבחן זה, שנחשב לחזק יותר מהמבחן הקודם, קיבלנו P-Value קטן מאוד ולכן נדחה את השערת האפס ברמת מובהקות 5% ונאמר שהנחת הנורמאליות לא מתקיימת במודל זה.

```
> # Shapiro Wilk - Normality Test :
> shapiro.test(datasetNewX$stan_residuals)

        Shapiro-Wilk normality test

data:  datasetNewX$stan_residuals
W = 0.77304, p-value = 1.266e-12
```

בדיקת הנחת הלינאריות:

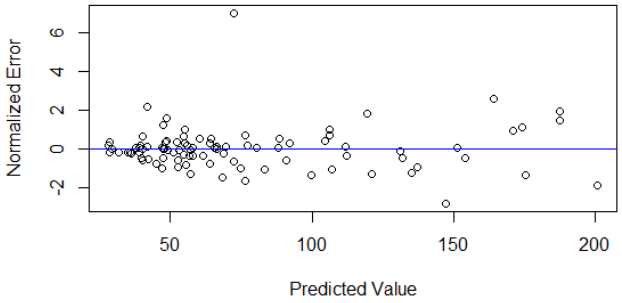
בהסתכלות על תרשים פיזור השגיאות המתוקננות לעומת ערך החיזוי ניתן לראות פיזור יחסית אחיד סביב האפס, לכן נסיק כי הנחת הלינאריות מתקיימת ונבחן זאת באמצעות מבחן סטטיסטי Chow.

```
> # Chow - Linearity Test :
> sctest(FM, type = "Chow")

        M-fluctuation test

data:  FM
f(efp) = 1.4117, p-value = 0.4942
```



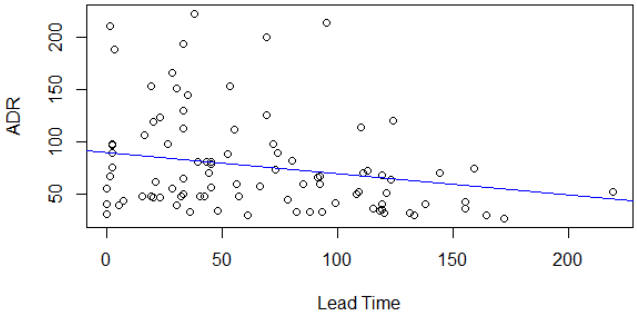
מבחן Chow לבחינת הנחת הלינאריות:

בהרצת מבחן Chow קיבלנו P-Value = 0.4942 ולכן נחליט שלא לדחות את השערת האפס ברמת מובהקות 5% ונאמר שהנחת הלינאריות מתקיימת. תוצאות המבחן מחזקות את מה שראינו מהגרף.



בדיקת הנחת שוויון שונויות:

נבדוק את הנחת שוויון השונויות באמצעות תרשים פיזור השגיאות המתוקננות לעומת ערך החיזוי. על פי הגרף ניתן לראות שעבור ערכים גבוהים יותר מתקבלת שונות שגיאות גבוהה יותר. קיבלנו מעין צורה של משפך הגדל מצד שמאל לימין, לכן על פי הגרף נאמר ששוויון השונויות לא מתקיים במודל זה.



מבחן Goldfeld - Quandt לבחינת הנחת שוויון השונויות:

לצורך ביצוע המבחן חיפשנו משתנה במודל שחשדנו בו שיתכן שגורם לאי שוויון השונויות ומצאנו את המשתנה LeadTime כחשוד.

בהסתכלות על הגרף ניתן לראות פיזור בצורת משפך הקטן משמאל לימין ומראה על אי שוויון שונויות.

ביצענו בדיקה של השונויות בתחומים שונים למשתנה זה וקיבלנו:

```
> Vlt50<-var(datasetNewX$adr[datasetNewX$lead_time < 50])%>%print()
[1] 2248.032
> Vlt100<-var(datasetNewX$adr[datasetNewX$lead_time < 100 & datasetNewX$lead_time > 50])%>%print()
[1] 2480.621
> Vlt150<-var(datasetNewX$adr[datasetNewX$lead_time < 150 & datasetNewX$lead_time > 100])%>%print()
[1] 548.7154
> Vlt200<-var(datasetNewX$adr[datasetNewX$lead_time < 200 & datasetNewX$lead_time > 150])%>%print()
[1] 370.4317
```

התוצאות שקיבלנו מתיישבות עם הגרף.

נבצע מבחן Goldfeld - Quandt :

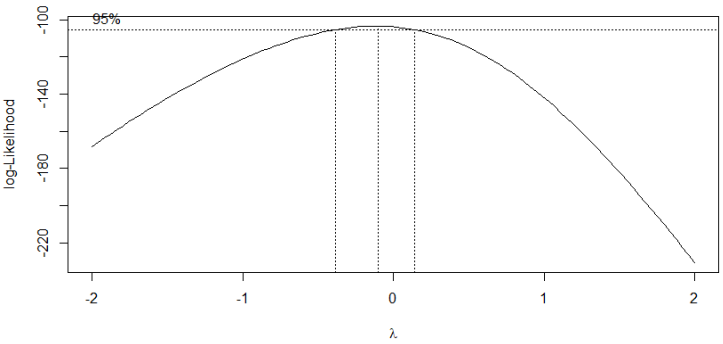
מתוצאות המבחן קיבלנו P-Value נמוך מאוד ולכן נדחה את השערת האפס ברמת מובהקות 5% ונאמר אין במודל שוויון שונויות.

```
> # Goldfeld Quandt - Test :
> gqtest(FM,alternative = "two.sided")

        Goldfeld-Quandt test

data:  FM
GQ = 2.5168, df1 = 45, df2 = 44, p-value = 0.002669
alternative hypothesis: variance changes from segment 1 to 2
```

4. שיפור המודל:

בבדיקות שביצענו על המודל קיבלנו שהנחת שוויון השונויות והנחת הנורמאליות לא מתקיימות. נבצע מבחן BoxCox כדי לבחון איזה טרנספורמציה לבצע על המשתנה המוסבר (adr). מהמבחן ניתן לראות שאפס נמצא ברווח הסמך ומכאן נבחר לבצע טרנספורמציה עבור $\lambda = 0$, כלומר ln(Y) .

$$y(\lambda) = \begin{cases} (y^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$



לאחר בדיקה עבור מודל ה ln(Y) ,
נראה שבעבור מודל זה הנחות שוויון שונויות וליניאריות כן מתקיימות, אך הנחת נורמליות השגיאות אינו מתקיים.
בדקנו מודלים נוספים וראינו למרות שיש מודלים בעלי R^2adj גבוהה יותר ממודל הלוג, פחות הנחות מתקיימות במודלים האחרים ולכן נבחר להישאר עם מודל זה.

טרנספורמציה	R^2adj	נורמליות	ליניאריות	שוויון שונויות
Y	0.7816	לא מתקיים	מתקיים	לא מתקיים
ln(Y)	0.7881	לא מתקיים	מתקיים	מתקיים
Y^2	0.706	לא מתקיים	מתקיים	לא מתקיים
sqrt(Y)	0.7962	לא מתקיים	מתקיים	לא מתקיים
$Y^{-0.5}$	0.7598	לא מתקיים	לא מתקיים	מתקיים

לאחר שבחרנו במודל ln(Y), נרצה לבדוק האם החלפה של המשתנים המסבירים במשתנים לאחר טרנספורמציה על המשתנים המסבירים יכולה לפתור לנו את בעיית ההנחות הדרושות למודל.
נבדוק 2 אופציות, טרנספורמציה לוגריתמית פולינומיאלית מסדר שני.

טרנספורמציה	R^2adj	נורמליות	ליניאריות	שוויון שונויות
Log	0.7689	לא מתקיים	מתקיים	לא מתקיים
X^2	0.7883	לא מתקיים	מתקיים	מתקיים

נראה שבעבור המודל עם הטרנספורמציה פולינומיאלית מסדר שני אנחנו מקבלים שהנחת הליניאריות ושוויון השונויות מתקיים לעומת המודל עם הטרנספורמציה הלוגריתמית, בנוסף נוכל לראות שמדד ה R^2adj גדל מעט ולכן אנחנו נבחר במודל זה להיות המודל הסופי שלי.

המודל הסופי:

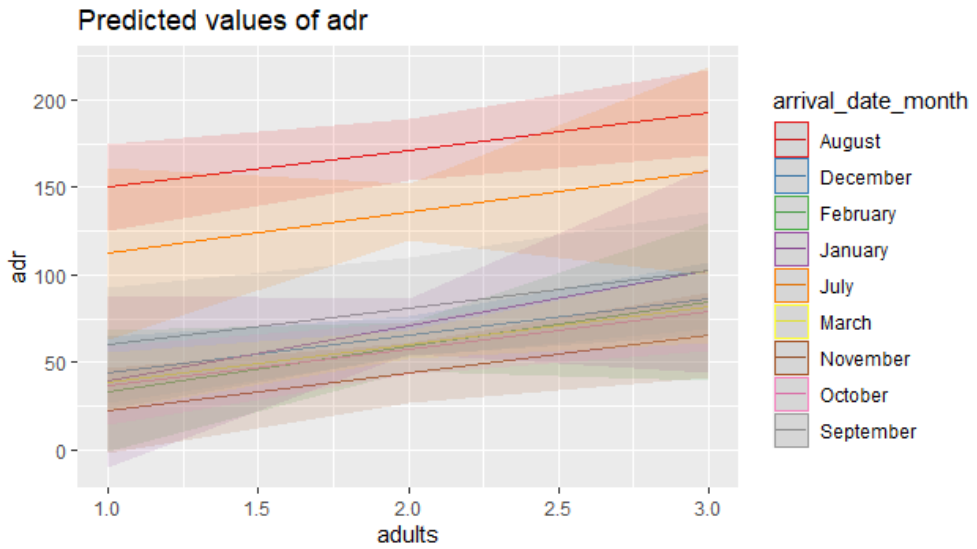
```
Log(adr)  ~ (lead_time)^2  + (adults)^2 +
            factor(arrival_date_month)  +
            factor(country)   +
            factor(reserved_room_type)  +
            factor(meal)
```

הבדיקות בנספחים.
צריך פה קצת בלבול שכל.

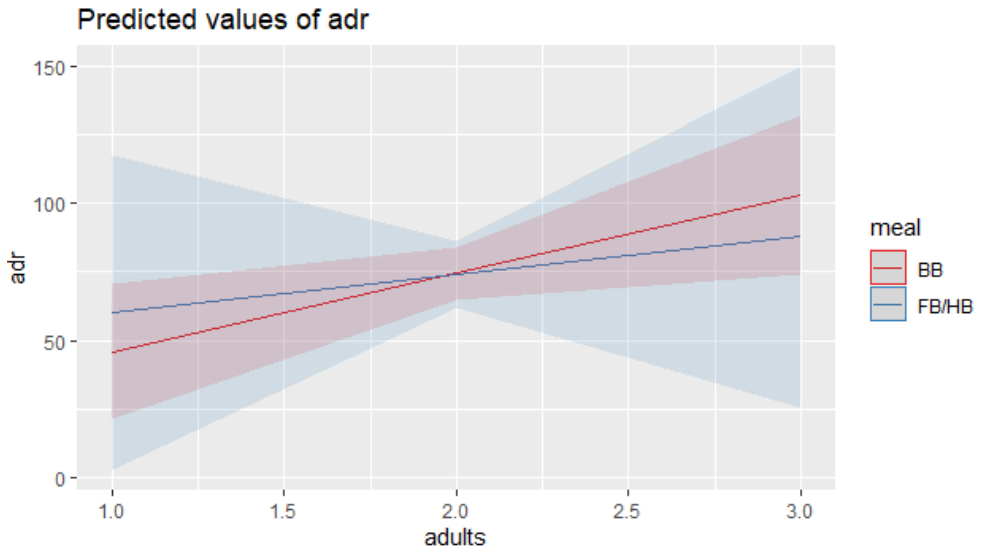
נספחים

אינטראקציות נוספות שנבדקו :

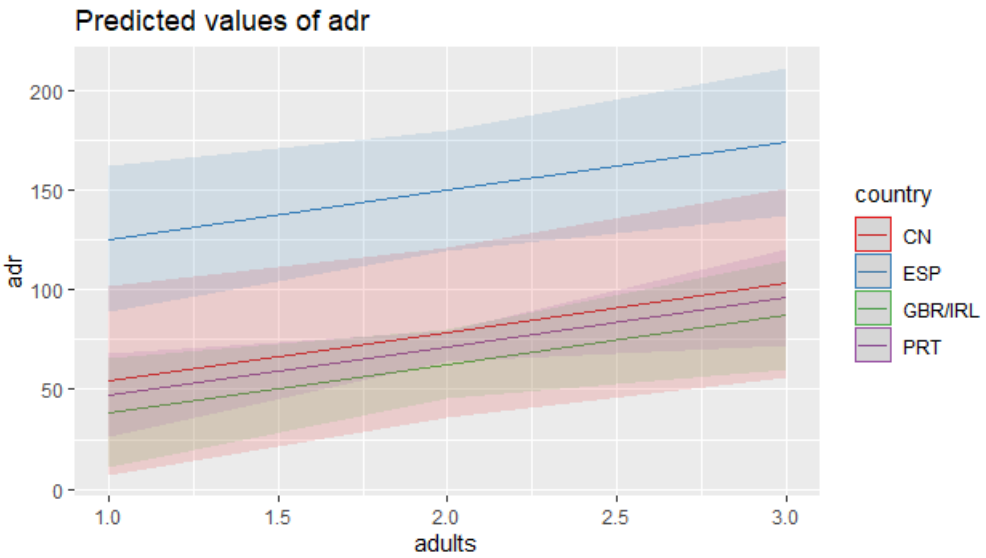
.X5 - Number of Adults & X2 - Months



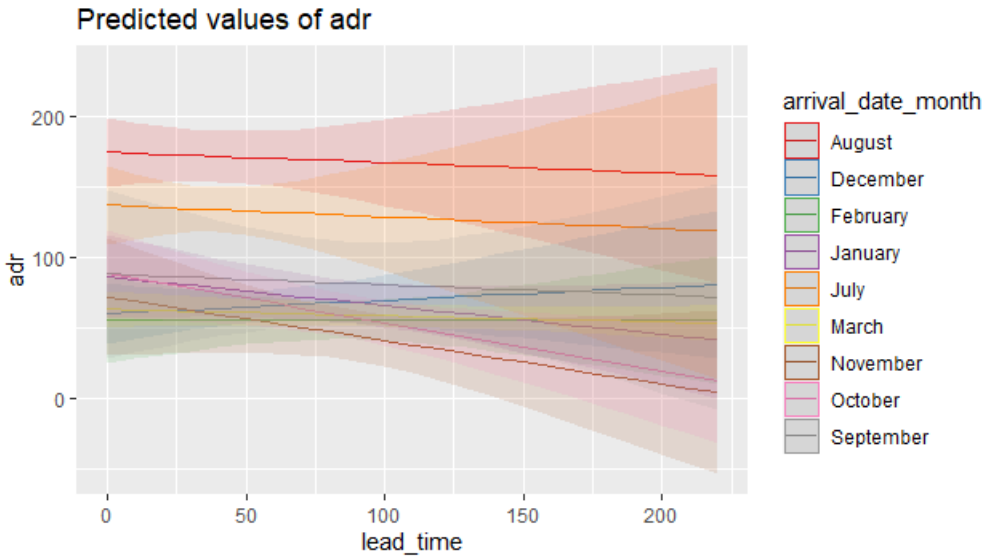
.X5 - Number of Adults & X6 - Meal



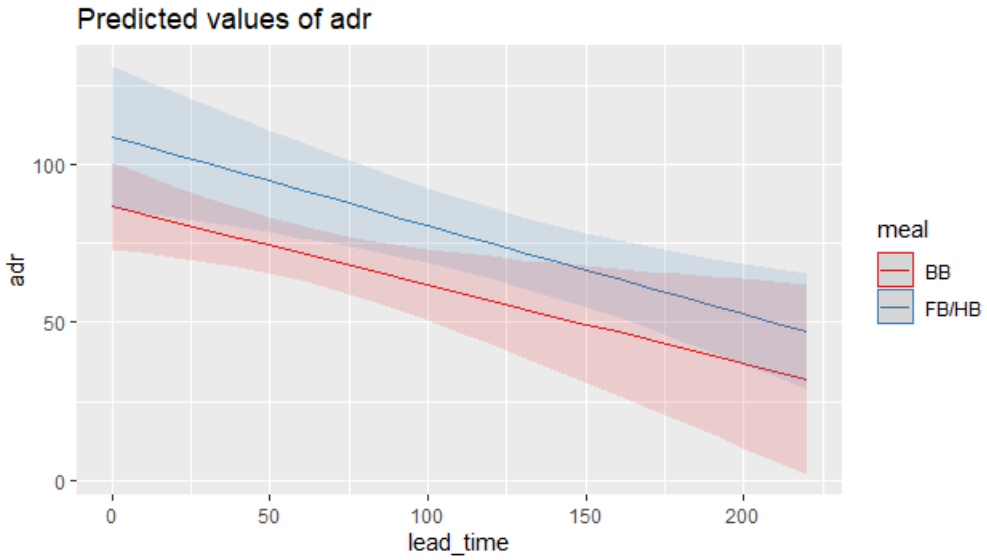
.X5 - Number of Adults & X7 - Countries



.X1 - Lead Time & X2 - Months



.X1 - Lead Time & X6 - Meal



בניית המודל :

Full Model

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.48 on 104 degrees of freedom
Multiple R-squared:  0.8175,	Adjusted R-squared:  0.7825 
F-statistic:  23.3 on 20 and 104 DF,  p-value: < 2.2e-16

> FRadj<-summary(FMnew)$adj.r.squared
> cat("Full Model - AIC:",FAIC[2])%>%
+   cat("\nFull Model - BIC:",FBIC[2])%>%
+   cat("\nFull Model - Radj:",FRadj)
Full Model - AIC: 761.3804
Full Model - BIC: 820.775
Full Model - Radj: 0.7824574
```

Empty Model

```
> cat("Empty Model - AIC:",EAIC)%>%
+   cat("\nEmpty Model - BIC:",EBIC)
Empty Model - AIC: 934.0368
Empty Model - BIC: 936.8651
>
```

AIC – FORWARD

```
> # AIC Forward Regression
> fwd.Model.AIC <-  step(Emp, direction='forward',scope=formula(FMnew))
Start:  AIC=934.04
adr ~ 1

                                        Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$arrival_date_month) 8    138777  77563 821.82
+ factor(datasetNewX$reserved_room_type) 3     44002 172339 911.61
+ datasetNewX$reserved_room_type         3     44002 172339 911.61
+ factor(datasetNewX$country)            3     38832 177508 915.31
+ datasetNewX$lead_time                  1     22646 193695 922.22
+ datasetNewX$adults                     1      8562 207778 930.99
<none>                                               216340 934.04
+ factor(datasetNewX$meal)               1        63 216277 936.00

Step:  AIC=821.82
adr ~ factor(datasetNewX$arrival_date_month)

                                        Df Sum of Sq   RSS    AIC
+ factor(datasetNewX$reserved_room_type) 3   19846.1 57717 790.87
+ datasetNewX$reserved_room_type         3   19846.1 57717 790.87
+ datasetNewX$adults                     1    6256.2 71307 813.30
+ factor(datasetNewX$country)            3    5756.4 71807 818.18
+ factor(datasetNewX$meal)               1    3159.2 74404 818.62
+ datasetNewX$lead_time                  1    2458.5 75105 819.79
<none>                                             77563 821.82

Step:  AIC=790.87
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type)

                            Df Sum of Sq   RSS    AIC
+ factor(datasetNewX$country) 3    9295.4 48422 774.92
+ factor(datasetNewX$meal)    1    7227.8 50489 776.15
+ datasetNewX$adults          1    2061.9 55655 788.33
<none>                                    57717 790.87
+ datasetNewX$lead_time       1      61.1 57656 792.74

Step:  AIC=774.92
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$country)

                           Df Sum of Sq   RSS    AIC
+ factor(datasetNewX$meal) 1    4834.9 43587 763.77
+ datasetNewX$adults       1    1601.7 46820 772.72
<none>                                  48422 774.92
+ datasetNewX$lead_time    1       2.1 48419 776.92
```

```
Step:  AIC=763.77
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$country) + factor(datasetNewX$meal)

                        Df Sum of Sq   RSS    AIC
+ datasetNewX$lead_time  1   1380.42 42206 761.75
+ datasetNewX$adults     1    713.79 42873 763.71
<none>                               43587 763.77

Step:  AIC=761.75
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$country) + factor(datasetNewX$meal) +
    datasetNewX$lead_time

                      Df Sum of Sq   RSS    AIC
+ datasetNewX$adults  1    1429.5 40777 759.44
<none>                            42206 761.75

Step:  AIC=759.44
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$country) + factor(datasetNewX$meal) +
    datasetNewX$lead_time + datasetNewX$adults

       Df Sum of Sq   RSS    AIC
<none>              40777 759.44
>
```

```
fwd.Model.AIC - AIC: 759.4443
fwd.Model.AIC - BIC: 810.354
fwd.Model.AIC - Radj: 0.7815695
```

BIC – FORWD

```
> # BIC Forward Regression
> fwd.Model.BIC <-  step(Emp, direction='forward',scope=formula(FMnew), k = log(nrow(datasetNewX)))
Start:  AIC=936.87
adr ~ 1

                                        Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$arrival_date_month)  8    138777  77563 847.27
+ factor(datasetNewX$reserved_room_type)  3     44002 172339 922.93
+ datasetNewX$reserved_room_type          3     44002 172339 922.93
+ factor(datasetNewX$country)             3     38832 177508 926.62
+ datasetNewX$lead_time                   1     22646 193695 927.87
+ datasetNewX$adults                      1      8562 207778 936.65
<none>                                                216340 936.87
+ factor(datasetNewX$meal)                1        63 216277 941.66

Step:  AIC=847.27
adr ~ factor(datasetNewX$arrival_date_month)

                                        Df Sum of Sq   RSS    AIC
+ factor(datasetNewX$reserved_room_type)  3   19846.1 57717 824.81
+ datasetNewX$reserved_room_type          3   19846.1 57717 824.81
+ datasetNewX$adults                      1    6256.2 71307 841.59
+ factor(datasetNewX$meal)                1    3159.2 74404 846.90
<none>                                                77563 847.27
+ datasetNewX$lead_time                   1    2458.5 75105 848.07
+ factor(datasetNewX$country)             3    5756.4 71807 852.12

Step:  AIC=824.81
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type)

                              Df Sum of Sq   RSS    AIC
+ factor(datasetNewX$meal)     1    7227.8 50489 812.92
+ factor(datasetNewX$country)  3    9295.4 48422 817.35
<none>                                     57717 824.81
+ datasetNewX$adults           1    2061.9 55655 825.10
+ datasetNewX$lead_time        1      61.1 57656 829.51

Step:  AIC=812.92
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$meal)

                              Df Sum of Sq   RSS    AIC
+ factor(datasetNewX$country)  3    6902.6 43587 809.03
+ datasetNewX$lead_time        1    2393.4 48096 811.68
<none>                                     50489 812.92
+ datasetNewX$adults           1     773.7 49716 815.82

Step:  AIC=809.03
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$meal) + factor(datasetNewX$country)

                        Df Sum of Sq   RSS    AIC
<none>                               43587 809.03
+ datasetNewX$lead_time  1   1380.42 42206 809.83
+ datasetNewX$adults     1    713.79 42873 811.79
>
```

```
fwd.Model.BIC - AIC: 763.7742
fwd.Model.BIC - BIC: 809.0272
fwd.Model.BIC - Radj: 0.7708017
```

AIC – BACK

```
> # AIC Backward Regression
> bw.Model.AIC <-  step(FMnew, direction='backward',scope=~1)
Start:  AIC=761.38
datasetNewX$adr ~ datasetNewX$lead_time + datasetNewX$adults +
    factor(datasetNewX$arrival_date_month) + factor(datasetNewX$country) +
    factor(datasetNewX$reserved_room_type) + factor(datasetNewX$meal) +
    datasetNewX$lead_time * datasetNewX$reserved_room_type


Step:  AIC=761.38
datasetNewX$adr ~ datasetNewX$lead_time + datasetNewX$adults +
    factor(datasetNewX$arrival_date_month) + factor(datasetNewX$country) +
    factor(datasetNewX$meal) + datasetNewX$reserved_room_type +
    datasetNewX$lead_time:datasetNewX$reserved_room_type

                                                Df Sum of Sq    RSS    AIC
- datasetNewX$lead_time:datasetNewX$reserved_room_type  3      1304  40777 759.44
<none>                                                          39472 761.38
- datasetNewX$adults                                 1      1217  40690 763.18
- factor(datasetNewX$country)                        3      5789  45262 772.49
- factor(datasetNewX$meal)                           1      5650  45122 776.10
- factor(datasetNewX$arrival_date_month)             8     67717 107190 870.26

Step:  AIC=759.44
datasetNewX$adr ~ datasetNewX$lead_time + datasetNewX$adults +
    factor(datasetNewX$arrival_date_month) + factor(datasetNewX$country) +
    factor(datasetNewX$meal) + datasetNewX$reserved_room_type

                                        Df Sum of Sq    RSS    AIC
<none>                                                40777 759.44
- datasetNewX$adults                     1      1429  42206 761.75
- datasetNewX$lead_time                  1      2096  42873 763.71
- factor(datasetNewX$country)            3      5577  46354 769.47
- factor(datasetNewX$meal)               1      5817  46593 774.11
- datasetNewX$reserved_room_type         3     16093  56870 795.03
- factor(datasetNewX$arrival_date_month) 8     77102 117879 876.14
> bw.A.AIC<-extractAIC(bw.Model.AIC)[2]
```

```
bw.Model.AIC - AIC: 759.4443
bw.Model.AIC - BIC: 810.354
bw.Model.AIC - Radj: 0.7815695
```

BIC – BACK

```
> # BIC Backward Regression
> bw.Model.BIC <-  step(FMnew, direction='backward',scope=~1, k = log(nrow(datasetNewX)))
Start:  AIC=820.78
datasetNewX$adr ~ datasetNewX$lead_time + datasetNewX$adults +
    factor(datasetNewX$arrival_date_month) + factor(datasetNewX$country) +
    factor(datasetNewX$reserved_room_type) + factor(datasetNewX$meal) +
    datasetNewX$lead_time * datasetNewX$reserved_room_type


Step:  AIC=820.78
datasetNewX$adr ~ datasetNewX$lead_time + datasetNewX$adults +
    factor(datasetNewX$arrival_date_month) + factor(datasetNewX$country) +
    factor(datasetNewX$meal) + datasetNewX$reserved_room_type +
    datasetNewX$lead_time:datasetNewX$reserved_room_type

                                                   Df Sum of Sq    RSS    AIC
- datasetNewX$lead_time:datasetNewX$reserved_room_type  3      1304  40777 810.35
- datasetNewX$adults                                   1      1217  40690 819.74
<none>                                                              39472 820.78
- factor(datasetNewX$country)                          3      5789  45262 823.40
- factor(datasetNewX$meal)                             1      5650  45122 832.67
- factor(datasetNewX$arrival_date_month)               8     67717 107190 907.02

Step:  AIC=810.35
datasetNewX$adr ~ datasetNewX$lead_time + datasetNewX$adults +
    factor(datasetNewX$arrival_date_month) + factor(datasetNewX$country) +
    factor(datasetNewX$meal) + datasetNewX$reserved_room_type

                                        Df Sum of Sq    RSS    AIC
- datasetNewX$adults                     1      1429  42206 809.83
<none>                                                40777 810.35
- datasetNewX$lead_time                  1      2096  42873 811.79
- factor(datasetNewX$country)            3      5577  46354 811.89
- factor(datasetNewX$meal)               1      5817  46593 822.19
- datasetNewX$reserved_room_type         3     16093  56870 837.45
- factor(datasetNewX$arrival_date_month) 8     77102 117879 904.42

Step:  AIC=809.83
datasetNewX$adr ~ datasetNewX$lead_time + factor(datasetNewX$arrival_date_month) +
    factor(datasetNewX$country) + factor(datasetNewX$meal) +
    datasetNewX$reserved_room_type

                                        Df Sum of Sq    RSS    AIC
- datasetNewX$lead_time                  1      1380  43587 809.03
<none>                                                42206 809.83
- factor(datasetNewX$country)            3      5890  48096 811.68
- factor(datasetNewX$meal)               1      6213  48419 822.17
- datasetNewX$reserved_room_type         3     21700  63906 847.20
- factor(datasetNewX$arrival_date_month) 8     77838 120044 901.87

Step:  AIC=809.03
datasetNewX$adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$country) +
    factor(datasetNewX$meal) + datasetNewX$reserved_room_type

                                        Df Sum of Sq    RSS    AIC
<none>                                                43587 809.03
- factor(datasetNewX$country)            3      6903  50489 812.92
- factor(datasetNewX$meal)               1      4835  48422 817.35
- datasetNewX$reserved_room_type         3     26452  70039 853.83
- factor(datasetNewX$arrival_date_month) 8     82427 126014 903.11
```

```
bw.Model.BIC - AIC: 763.7742
bw.Model.BIC - BIC: 809.0272
bw.Model.BIC - Radj: 0.7708017
```

AIC – STEPS

```
> # AIC Steps Regression
> SW.AIC <- step(Emp,direction = "both",scope = formula(FMnew))
Start:  AIC=934.04
adr ~ 1

                                          Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$arrival_date_month)  8    138777  77563 821.82
+ factor(datasetNewX$reserved_room_type)  3     44002 172339 911.61
+ datasetNewX$reserved_room_type          3     44002 172339 911.61
+ factor(datasetNewX$country)             3     38832 177508 915.31
+ datasetNewX$lead_time                   1     22646 193695 922.22
+ datasetNewX$adults                      1      8562 207778 930.99
<none>                                                216340 934.04
+ factor(datasetNewX$meal)                1        63 216277 936.00

Step:  AIC=821.82
adr ~ factor(datasetNewX$arrival_date_month)

                                          Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$reserved_room_type)  3     19846  57717 790.87
+ datasetNewX$reserved_room_type          3     19846  57717 790.87
+ datasetNewX$adults                      1      6256  71307 813.30
+ factor(datasetNewX$country)             3      5756  71807 818.18
+ factor(datasetNewX$meal)                1      3159  74404 818.62
+ datasetNewX$lead_time                   1      2458  75105 819.79
<none>                                                77563 821.82
- factor(datasetNewX$arrival_date_month)  8    138777 216340 934.04

Step:  AIC=790.87
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type)

                                          Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$country)             3      9295  48422 774.92
+ factor(datasetNewX$meal)                1      7228  50489 776.15
+ datasetNewX$adults                      1      2062  55655 788.33
<none>                                                57717 790.87
+ datasetNewX$lead_time                   1        61  57656 792.74
- factor(datasetNewX$reserved_room_type)  3     19846  77563 821.82
- factor(datasetNewX$arrival_date_month)  8    114622 172339 911.61

Step:  AIC=774.92
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$country)

                                          Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$meal)                1      4835  43587 763.77
+ datasetNewX$adults                      1      1602  46820 772.72
<none>                                                48422 774.92
+ datasetNewX$lead_time                   1         2  48419 776.92
- factor(datasetNewX$country)             3      9295  57717 790.87
- factor(datasetNewX$reserved_room_type)  3     23385  71807 818.18
- factor(datasetNewX$arrival_date_month)  8     77912 126333 878.80
```

```
Step:  AIC=763.77
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$country) + factor(datasetNewX$meal)

                                          Df Sum of Sq    RSS    AIC
+ datasetNewX$lead_time                   1      1380  42206 761.75
+ datasetNewX$adults                      1       714  42873 763.71
<none>                                                43587 763.77
- factor(datasetNewX$meal)                1      4835  48422 774.92
- factor(datasetNewX$country)             3      6903  50489 776.15
- factor(datasetNewX$reserved_room_type)  3     26452  70039 817.06
- factor(datasetNewX$arrival_date_month)  8     82427 126014 880.48

Step:  AIC=761.75
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$country) + factor(datasetNewX$meal) +
    datasetNewX$lead_time

                                          Df Sum of Sq    RSS    AIC
+ datasetNewX$adults                      1      1429  40777 759.44
<none>                                                42206 761.75
- datasetNewX$lead_time                   1      1380  43587 763.77
- factor(datasetNewX$country)             3      5890  48096 772.08
- factor(datasetNewX$meal)                1      6213  48419 776.92
- factor(datasetNewX$reserved_room_type)  3     21700  63906 807.61
- factor(datasetNewX$arrival_date_month)  8     77838 120044 876.41

Step:  AIC=759.44
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$country) + factor(datasetNewX$meal) +
    datasetNewX$lead_time + datasetNewX$adults

                                          Df Sum of Sq    RSS    AIC
<none>                                                40777 759.44
- datasetNewX$adults                      1      1429  42206 761.75
- datasetNewX$lead_time                   1      2096  42873 763.71
- factor(datasetNewX$country)             3      5577  46354 769.47
- factor(datasetNewX$meal)                1      5817  46593 774.11
- factor(datasetNewX$reserved_room_type)  3     16093  56870 795.03
- factor(datasetNewX$arrival_date_month)  8     77102 117879 876.14
```

```
SW.AIC - AIC: 759.4443
SW.AIC - BIC: 810.354
SW.AIC - Radj: 0.7815695
```

BIC – STEPS

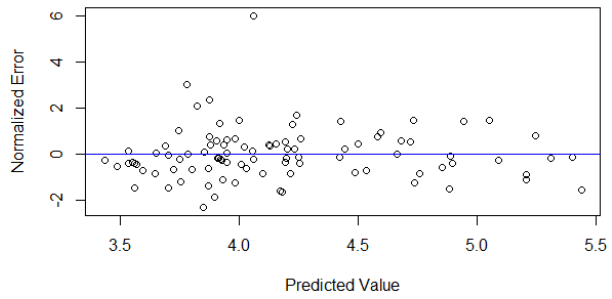
```
> # BIC Steps Regression
> SW.BIC <- step(Emp,direction = "both",scope = formula(FMnew), k = log(nrow(datasetN
Start:  AIC=936.87
adr ~ 1

                                        Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$arrival_date_month)  8    138777  77563 847.27
+ factor(datasetNewX$reserved_room_type)  3     44002 172339 922.93
+ datasetNewX$reserved_room_type          3     44002 172339 922.93
+ factor(datasetNewX$country)             3     38832 177508 926.62
+ datasetNewX$lead_time                   1     22646 193695 927.87
+ datasetNewX$adults                      1      8562 207778 936.65
<none>                                                216340 936.87
+ factor(datasetNewX$meal)                1        63 216277 941.66

Step:  AIC=847.27
adr ~ factor(datasetNewX$arrival_date_month)

                                        Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$reserved_room_type)  3     19846  57717 824.81
+ datasetNewX$reserved_room_type          3     19846  57717 824.81
+ datasetNewX$adults                      1      6256  71307 841.59
+ factor(datasetNewX$meal)                1      3159  74404 846.90
<none>                                                 77563 847.27
+ datasetNewX$lead_time                   1      2458  75105 848.07
+ factor(datasetNewX$country)             3      5756  71807 852.12
- factor(datasetNewX$arrival_date_month)  8    138777 216340 936.87

Step:  AIC=824.81
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type)

                                        Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$meal)                1      7228  50489 812.92
+ factor(datasetNewX$country)             3      9295  48422 817.35
<none>                                                 57717 824.81
+ datasetNewX$adults                      1      2062  55655 825.10
+ datasetNewX$lead_time                   1        61  57656 829.51
- factor(datasetNewX$reserved_room_type)  3     19846  77563 847.27
- factor(datasetNewX$arrival_date_month)  8    114622 172339 922.93
```

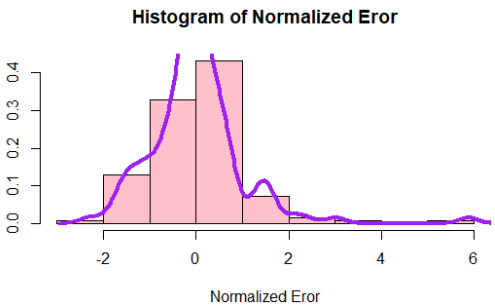
```
Step:  AIC=812.92
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$meal)

                                        Df Sum of Sq    RSS    AIC
+ factor(datasetNewX$country)             3      6903  43587 809.03
+ datasetNewX$lead_time                   1      2393  48096 811.68
<none>                                                 50489 812.92
+ datasetNewX$adults                      1       774  49716 815.82
- factor(datasetNewX$meal)                1      7228  57717 824.81
- factor(datasetNewX$reserved_room_type)  3     23915  74404 846.90
- factor(datasetNewX$arrival_date_month)  8    118936 169425 925.62

Step:  AIC=809.03
adr ~ factor(datasetNewX$arrival_date_month) + factor(datasetNewX$reserved_room_type) +
    factor(datasetNewX$meal) + factor(datasetNewX$country)

                                        Df Sum of Sq    RSS    AIC
<none>                                                 43587 809.03
+ datasetNewX$lead_time                   1      1380  42206 809.83
+ datasetNewX$adults                      1       714  42873 811.79
- factor(datasetNewX$country)             3      6903  50489 812.92
- factor(datasetNewX$meal)                1      4835  48422 817.35
- factor(datasetNewX$reserved_room_type)  3     26452  70039 853.83
- factor(datasetNewX$arrival_date_month)  8     82427 126014 903.11
```

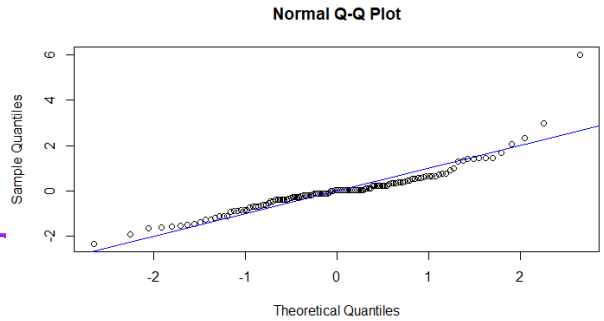
```
SW.BIC - AIC: 763.7742
SW.BIC - BIC: 809.0272
SW.BIC - Radj: 0.7708017
```

בדיקת הנחות מודלים לשיפור :

הנחות מודל עבור - Log(Y)








```
> # KS Test - Normality :
> ks.test(x = datasetNewX$stan_residuals, y = "pnorm",
+         alternative = "two.sided", exact = NULL)

        One-sample Kolmogorov-Smirnov test

data:  datasetNewX$stan_residuals
D = 0.12939, p-value = 0.03042
alternative hypothesis: two-sided
```

```
> # Shapiro Wilk - Normality Test :
> shapiro.test(datasetNewX$stan_residuals)

        Shapiro-Wilk normality test

data:  datasetNewX$stan_residuals
W = 0.87464, p-value = 7.192e-09
```

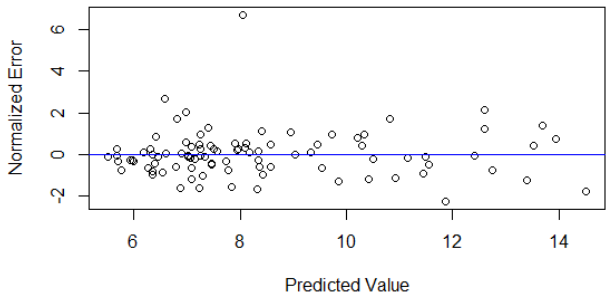
```
> # Goldfeld Quandt - Test :
> gqtest(FM_Log,alternative = "two.sided")

        Goldfeld-Quandt test

data:  FM_Log
GQ = 1.4563, df1 = 45, df2 = 44, p-value = 0.2146
alternative hypothesis: variance changes from segment 1 to 2
```

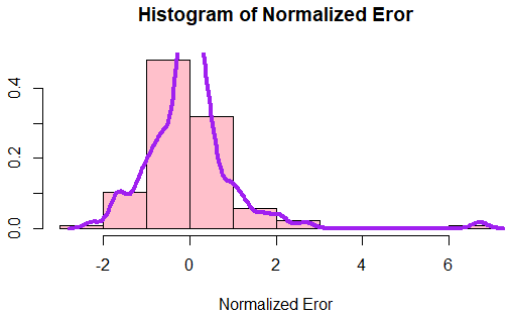
```
> # Chow - Linearity Test :
> sctest(FM_Log, type = "Chow")

        M-fluctuation test

data:  FM_Log
f(efp) = 1.7173, p-value = 0.09431
```

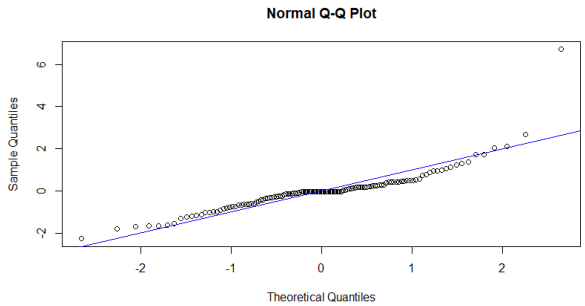
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2226 on 107 degrees of freedom
Multiple R-squared:  0.8172,	Adjusted R-squared:  0.7881
F-statistic: 28.14 on 17 and 107 DF,  p-value: < 2.2e-16
```

הנחת הליניאריות ושוויון השונויות מתקיימת.

הנחות המודל עבור - Sqrt(Y) Model








```
> # KS Test - Normality :
> ks.test(x = datasetNewX$stan_residuals, y = "pnorm",
+         alternative = "two.sided", exact = NULL)

        One-sample Kolmogorov-Smirnov test

data:  datasetNewX$stan_residuals
D = 0.15366, p-value = 0.005465
alternative hypothesis: two-sided
```

```
> # Shapiro Wilk - Normality Test :
> shapiro.test(datasetNewX$stan_residuals)

        Shapiro-Wilk normality test

data:  datasetNewX$stan_residuals
W = 0.81606, p-value = 3.298e-11
```

```
> # Goldfeld Quandt - Test :
> gqtest(FM_Sqrt,alternative = "two.sided")

        Goldfeld-Quandt test

data:  FM_Sqrt
GQ = 2.0898, df1 = 45, df2 = 44, p-value = 0.0158
alternative hypothesis: variance changes from segment 1 to 2
```

```
> # Chow - Linearity Test :
> sctest(FM_Sqrt, type = "Chow")

        M-fluctuation test

data:  FM_Sqrt
f(efp) = 1.5322, p-value = 0.2826
```

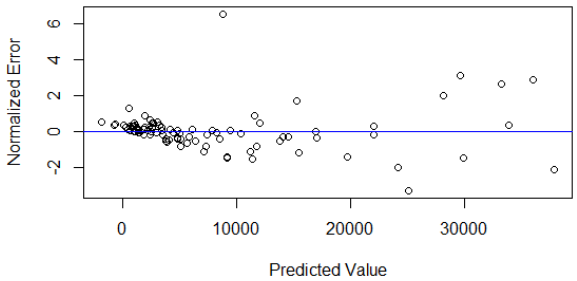
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9765 on 107 degrees of freedom
Multiple R-squared:  0.8241,	Adjusted R-squared:  0.7962
F-statistic:  29.5 on 17 and 107 DF,  p-value: < 2.2e-16
```

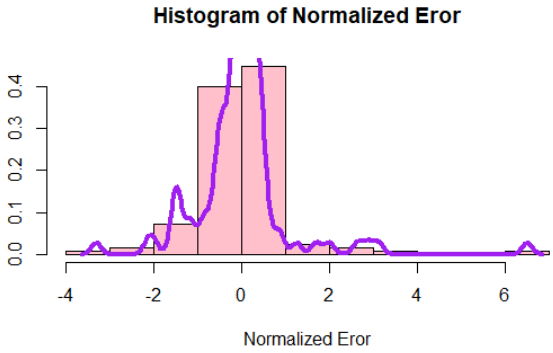
רק הנחת הליניאריות מתקיימת.

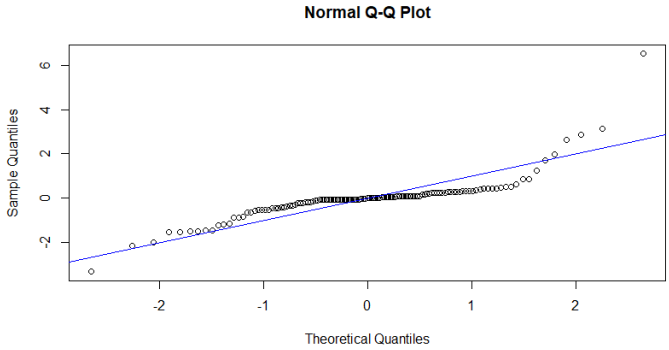
לאחר שנבצע טרנספורמציית שורש עבור המשתנים Lead Time ו Adults :

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9918 on 107 degrees of freedom
Multiple R-squared:  0.8186,	Adjusted R-squared:  0.7898
F-statistic:  28.4 on 17 and 107 DF,  p-value: < 2.2e-16
```

הנחות המודל עבור - Squared - (Y)^2








```
> # KS Test - Normality :
> ks.test(x = datasetNewX$stan_residuals, y = "pnorm",
+         alternative = "two.sided", exact = NULL)

        One-sample Kolmogorov-Smirnov test

data:  datasetNewX$stan_residuals
D = 0.22457, p-value = 6.692e-06
alternative hypothesis: two-sided
```

```
> # Shapiro Wilk - Normality Test :
> shapiro.test(datasetNewX$stan_residuals)

        Shapiro-Wilk normality test

data:  datasetNewX$stan_residuals
W = 0.73485, p-value = 9.861e-14
```

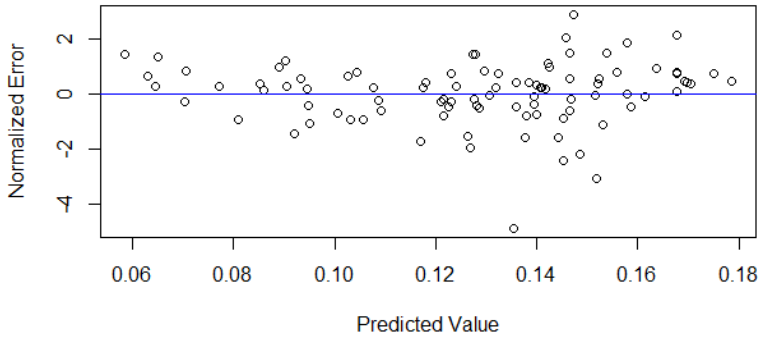
```
> # Goldfeld Quandt - Test :
> gqtest(FM_Squared,alternative = "two.sided")

        Goldfeld-Quandt test

data:  FM_Squared
GQ = 2.3169, df1 = 45, df2 = 44, p-value = 0.006103
alternative hypothesis: variance changes from segment 1 to 2
```

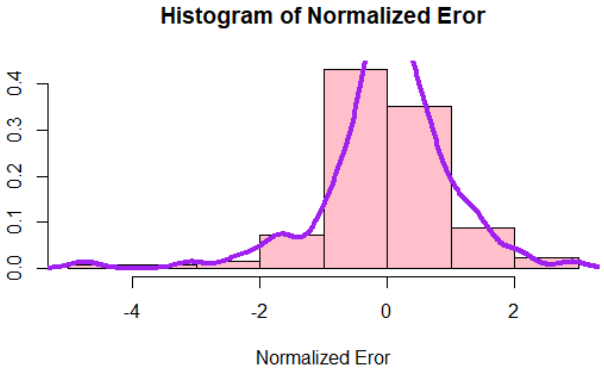
```
> # Chow - Linearity Test :
> sctest(FM_Squared, type = "Chow")

        M-fluctuation test

data:  FM_Squared
f(efp) = 1.5554, p-value = 0.2497
```

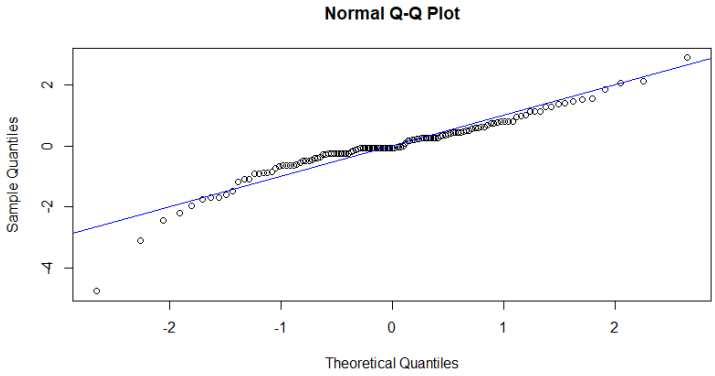
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5145 on 107 degrees of freedom
Multiple R-squared:  0.7463,	Adjusted R-squared:  0.706
F-statistic: 18.51 on 17 and 107 DF,  p-value: < 2.2e-16
```

רק הנחת הליניאריות מתקיימת.

הנחות המודל עבור – Model -0.5^(Y)








```
> # KS Test - Normality :
> ks.test(x = datasetNewX$stan_residuals, y = "pnorm",
+         alternative = "two.sided", exact = NULL)

        One-sample Kolmogorov-Smirnov test

data:  datasetNewX$stan_residuals
D = 0.12729, p-value = 0.03483
alternative hypothesis: two-sided
```

```
> # Shapiro Wilk - Normality Test :
> shapiro.test(datasetNewX$stan_residuals)

        Shapiro-Wilk normality test

data:  datasetNewX$stan_residuals
W = 0.93027, p-value = 6.759e-06
```

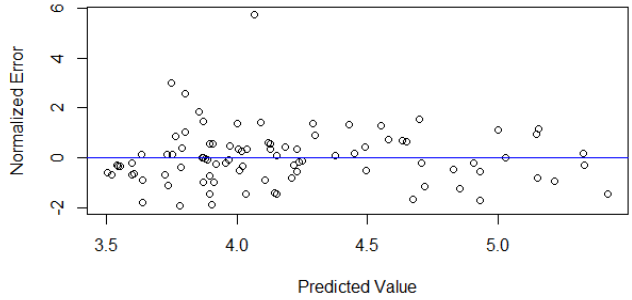
```
> # Goldfeld Quandt - Test :
> gqtest(FM_NegSqrt,alternative = "two.sided")

        Goldfeld-Quandt test

data:  FM_NegSqrt
GQ = 0.97678, df1 = 45, df2 = 44, p-value = 0.937
alternative hypothesis: variance changes from segment 1 to 2
```

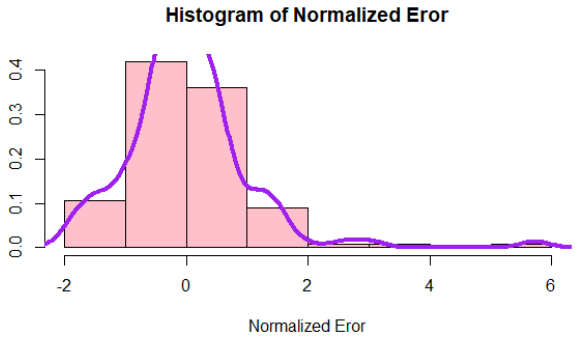
```
> # Chow - Linearity Test :
> sctest(FM_NegSqrt, type = "Chow")

        M-fluctuation test

data:  FM_NegSqrt
f(efp) = 1.9106, p-value = 0.02402
```

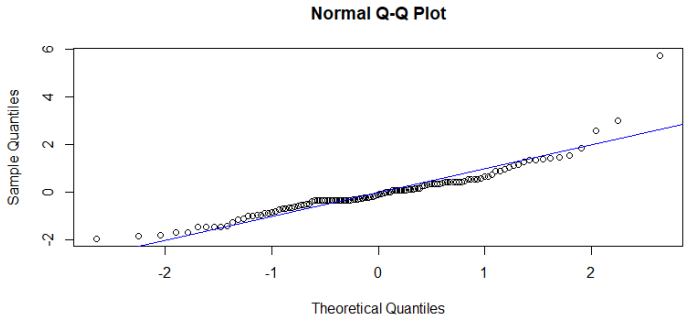
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01431 on 107 degrees of freedom
Multiple R-squared:  0.7928,	Adjusted R-squared:  0.7598
F-statistic: 24.08 on 17 and 107 DF,  p-value: < 2.2e-16
```

לא עומד בהנחות המודל מלבד שוויון שונויות.

בדיקת הנחות למודל Log עם טרנספורמציה למשתנים מסבירים ב Log :








```
> # KS Test - Normality :
> ks.test(x = datasetNewX$stan_residuals, y = "pnorm",
+         alternative = "two.sided", exact = NULL)

        One-sample Kolmogorov-Smirnov test

data:  datasetNewX$stan_residuals
D = 0.1165, p-value = 0.07291
alternative hypothesis: two-sided
```

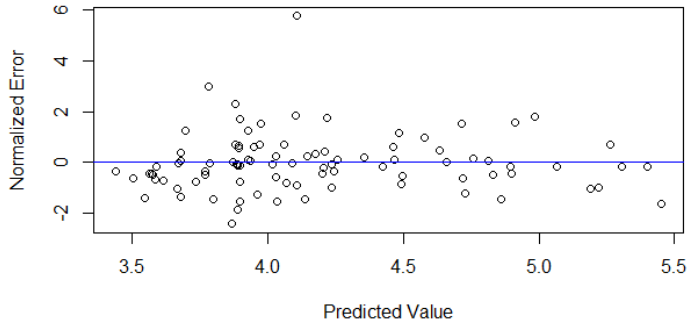
```
> # Shapiro Wilk - Normality Test :
> shapiro.test(datasetNewX$stan_residuals)

        Shapiro-Wilk normality test

data:  datasetNewX$stan_residuals
W = 0.88528, p-value = 3.042e-08
```

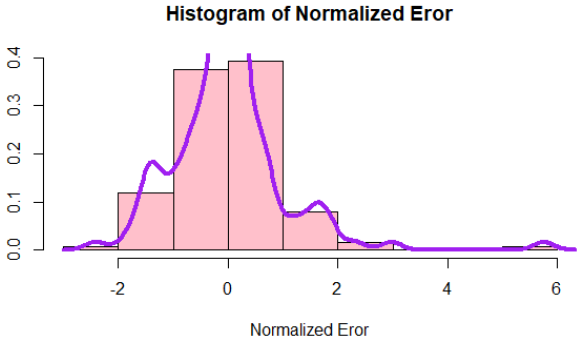
```
> # Goldfeld Quandt - Test :
> gqtest(FM_Log,alternative = "two.sided")

        Goldfeld-Quandt test

data:  FM_Log
GQ = 1.2744, df1 = 43, df2 = 43, p-value = 0.4298
alternative hypothesis: variance changes from segment 1 to 2
```

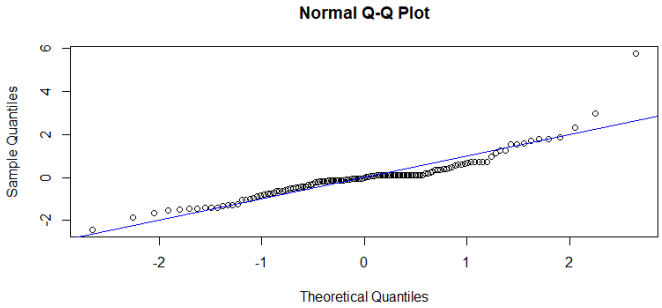
```
> # Chow - Linearity Test :
> sctest(FM_Log, type = "Chow")

        M-fluctuation test

data:  FM_Log
f(efp) = 2.2413, p-value = 0.001558

> # Goldfeld Quandt - Test :
> gqtest(FM_Log,alternative = "two.sided")
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.232 on 104 degrees of freedom
Multiple R-squared:  0.8014,	Adjusted R-squared:  0.7689
F-statistic: 24.68 on 17 and 104 DF,  p-value: < 2.2e-16
```

בדיקת הנחות למודל Log עם טרנספורמציה פולינומיאלית למשתנים מסבירים בסדר שני :








```
> # KS Test - Normality :
> ks.test(x = datasetNewX$stan_residuals, y = "pnorm",
+         alternative = "two.sided", exact = NULL)

        One-sample Kolmogorov-Smirnov test

data:  datasetNewX$stan_residuals
D = 0.17237, p-value = 0.001189
alternative hypothesis: two-sided
```

```
> # Shapiro Wilk - Normality Test :
> shapiro.test(datasetNewX$stan_residuals)

        Shapiro-Wilk normality test

data:  datasetNewX$stan_residuals
W = 0.87524, p-value = 7.656e-09
```

```
> # Goldfeld Quandt - Test :
> gqtest(FM_Log,alternative = "two.sided")

        Goldfeld-Quandt test

data:  FM_Log
GQ = 1.4485, df1 = 45, df2 = 44, p-value = 0.2212
alternative hypothesis: variance changes from segment 1 to 2
```

```
> # Chow - Linearity Test :
> sctest(FM_Log, type = "Chow")

        M-fluctuation test

data:  FM_Log
f(efp) = 1.6864, p-value = 0.1152
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2225 on 107 degrees of freedom
Multiple R-squared:  0.8173,    Adjusted R-squared:  0.7883
F-statistic: 28.16 on 17 and 107 DF,  p-value: < 2.2e-16
```