אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

המחלקה להנדסת תעשייה וניהול

קורס מודלים של רגרסיה לינארית

364.1.1061

תאריך הגשה: 24.1.2020

פרויקט בניתוח סטטיסטי של מאגרי נתונים טבלאיים- חלק ב'

קבוצה 45

מגישים:

שיר רוזנס 311119929

דור אלדר 312463839

עומרי זאבי 313327041

מרצה:

ד"ר הלל בר גרא

1. תוכן עניינים

2. תקציר מנהלים

במהלך הפרויקט הוקם מאגר נתונים המכיל נתונים על 100 שחקני כדורגל המשחקים ב-5 הליגות הבכירות באירופה, ועל בסיסו נבנה מודל רגרסיה מרובה. המודל המתואר בפרויקט נועד לאמוד את השכר השנתי של שחקן כדורגל (במיליוני יורו) וזאת כתלות בעשרה משתנים מסבירים שונים אשר נבחרו בקפידה. את הנתונים על עשרת משתנים המסבירים מצאנו באתרי האינטרנט www.whoscored.com/Statistics ו-www.capology.com/club אשר סיפקו את המידע הרלוונטי של כל משתנה מסביר עבור כל שחקן. השערתנו הייתה כי משתנים אלו יסבירו באופן הטוב ביותר את המשתנה המוסבר-השכר השנתי של שחקן כדורגל.

ראשית, בוצע ניתוח של הקשר התיאורטי בין כל אחד מהמשתנים המסבירים עם המשתנה המוסבר. שנית, בוצע ניתוח של כל משתנה מסביר עם שאר המשתנים המסבירים, ולאחר מכן בוצע ניתוח תיאורי של כל אחד מהמשתנים במודל. נבדקו גם חריגות ותרשימי פיזור בין זוגות משתנים כדי לזהות מגמתיות והשפעות הדדיות. בחלק זה בחנו תחילה את הקשרים הלינאריים בין כל אחד מהמשתנים המסבירים לבין המשתנה המוסבר באמצעות מקדם מתאם פירסון, כאשר משתנים שקיבלו ערך נמוך סומנו כפוטנציאליים להסרה מהמודל. זאת מכיוון שהתגלה מתאם הקרוב בערך מוחלט ל-0, מה שמעיד על כך שהוא חלש עבור משתנים אלו. בנוסף, מהתבוננות בתרשימי הפיזור של משתנים אלו כפונקציה של המשתנה המוסבר, החלטנו להסיר אותם מהמודל עקב חוסר ההשפעה של המשתנים עליו. הוחלט על 7 משתנים אשר על סמך בדיקות אלו הוצאו לחלוטין ממודל הרגרסיה המרובה.

שנית, הגדרנו משתני דמה ומשתני אינטראקציה עבור המודל שבאמצעותם בנינו מודל ראשוני וניסיוני לפרויקט. תהליך בחירת המשתנה המוסבר עבור משתנה האינטראקציה בוצע על סמך התרומה הגדולה ביותר להסברת המשתנה המוסבר. על מנת לבחור את המודל הסופי (בו התעסקנו בהמשך הפרויקט), לקחנו את המודל הראשוני וביצענו עליו מספר בדיקות: רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים. לאחר בדיקת מדדי כל אחד מהמודלים שהתקבלו מהאלגוריתמים, בחרנו את המודל הטוב ביותר על פי מדד BIC,AIC ועל פי מדד R^2_{adj}. המודל שהתקבל הוא:

$$\hat{y} = \beta_0 + \beta_1 x_5 + \beta_2 x_6 + \beta_3 CT_2 - \beta_4 x_5 * CT_2 - \beta_5 x_6 * CT_2$$

בהינתן המודל הראשוני הנ"ל, נבחנו הנחות המודל: ליניאריות, שוויון שונויות ונורמליות השגיאות. בדיקת ההנחות נעשתה בעזרת תרשימים ובעזרת מבחן סטטיסטי KS לבחינת הנחת הנורמליות, ונמצא כי הנחת שוויון השונויות אינה מתקיימת.

לשם תיקון המודל, ביצענו טרנספורמציה על המשתנה המוסבר לפי λ שהתקבל בתרשים cox-box וקיבלנו מודל חדש משופר אשר מקיים את הנחת שוויון השונויות.
לאחר התיקון התקבל המודל הבא:

$$\hat{y}^{0.25} = \beta_0 + \beta_1 x_5 + \beta_2 x_6 + \beta_3 CT_2 - \beta_4 x_5 * CT_2 - \beta_5 x_6 * CT_2$$

3. עיבוד מקדים

3.1 הגדרת משתנים

שם משתנה מסביר	יחידות מידה	רציף/בדיד/קטגוריאלי	תחום ערכים	פירוט
משכורת השחקן	מיליוני יורו	רציף	0.045-70.7	משכורת השנתית של שחקן במיליוני יורו
תפקיד	מילולי	קטגוריאלי	1-3	תפקידו של שחקן על המגרש בקבוצתו
קבוצה	מילולי	קטגוריאלי	1-2	הקבוצה הנוכחית בה משחק השחקן
נבחרת	מילולי	קטגורי אלי	1-4	הנבחרת הארצית לה שייך השחקן
גיל השחקן	שנים	רציף	19-34	גילו של השחקן
דירוג שחקן	ערך דירוג	רציף	7.22-8.75	מדד אשר כולל בתוכו מספר אלמנטים משוכללים אשר מצביעים על הנתונים הפיזיים והיכולות הטכניות של שחקן
ממוצע גולים למשחק AGPG	מס' גולים	רציף	0-1.2	ממוצע הגולים אשר הפקיע במשחק
ממוצע בישולים למשחק AAPG	מס' בישולים	רציף	0-0.73	ממוצע הפעמים שהשחקן מוסר לשחקן אחר מקבוצתו את הכדור והשחקן השני מבקיע גול מיד לאחר מסירה זו
מספר משחקים APP	מס' משחקים	רציף	9-16	מספר המשחקים בו השחקן שיחק במהלך עונה
אחוז מסירות מוצלחות למשחק PS	מס מסירות	רציף	53.7-92.8	מסירה של שחקן נספרת על ידי העברה של כדור משחקן בצורה ישירה לשחקן אחר מאותה קבוצה בלי שנלקח על ידי שחקן יריב
איבודי כדור למשחק DISP	מס' איבודי כדור	רציף	0-3	איבוד כדור על ידי שחקן נספר כאשר שחקן מעביר כדור לשחקן אחר מקבוצתו, ושחקן מקבוצה יריבה לוקח את הכדור במהלך מסירה זו.

פירוט משתנים קטגוריאליים:

תפקיד: 2-תפקיד התקפה 2-תפקיד קישור 3- תפקיד הגנה

קבוצה: 1- קבוצה בדרג בכירה 2- קבוצה בדרג נמוך

נבחרת: 1- אירופה 2- אמריקה 3- אפריקה 4-אסיה

3.2 הסרה של משתנים

בכדי לבדוק אילו משתנים נסיר מהמודל, נבדוק את המתאם בין המשתנה המוסבר לבין כל אחד מהמשתנים המסבירים. נשתמש במתאם פירסון (ρ) ותרשימי פיזור.

מתאם פירסון הינו מדד שימושי עבור סטטיסטיקה הסקתית שמטרתו הינה למצוא קשר סטטיסטי בין שני משתנים. ערך המתאם ינוע בין מינוס אחד לאחד ($-1 < \rho < 1$). התחום השלילי מסמן קשר שלילי בין המשתנים והתחום החיובי מסמן קשר חיובי בין המשתנים. כאמור, ככל שמקדם המתאם יהיה קרוב יותר ל -0 (בערך מוחלט), כך המתאם בין המשתנה המוסבר למשתנה המסביר יהיה נמוך יותר ובהתאם לכך נשקול להוציאו מהמודל. מקדם המתאם של Pearson בין שני משתנים מקריים מוגדר על ידי הנוסחה –

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

סוג הקשר במתאם Pearson הינו קשר ליניארי (המוגדר באמצעות קו ישר). משום שמדד זה אינו רלוונטי עבור המשתנים הקטגוריאליים, לא נבדוק אותם בסעיף זה. עבור כל משתנה (שאינו קטגוריאלי) נראה את המתאם שלו עם המשתנה המסביר (קוד ופלט בנספח 3.2.1). בנוסף, הצגנו בחלק א' תרשימי פיזור בין המשתנים. תרשימי פיזור אלו משמשים להצגת נתונים בעלי מאפיינים משותפים ולהצגת מגמות. בעזרת scatter plots נוח לבחון קורלציות בין שני משתנים. במידה ולא תהיה התאמה בין שני המשתנים, הגרף ייראה כמו אוסף נקודות אקראיות.

להלן טבלה המסכמת את מתאם פירסון עבור כל משתנה מסביר (שאינו קטגוריאלי) ותרשימי הפיזור לכל משתנה מסביר ביחס למשתנה המוסבר:

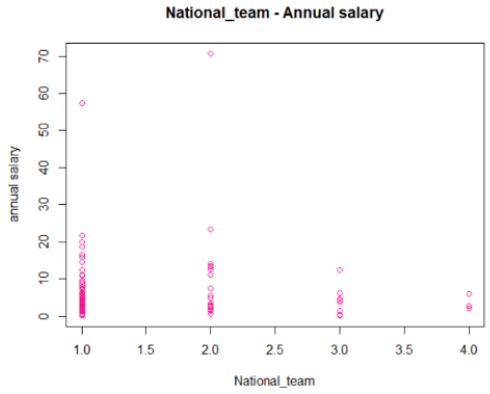
שם משתנה מסביר	מתאם פירסון
גיל השחקן	0.3802837
דירוג שחקן	0.5289814
ממוצע גולים למשחק AGPG	0.3859991
ממוצע בישולים למשחק AAPG	0.1040835
אחוז מסירות מוצלחות למשחק PS	0.07776205
איבודי כדור למשחק DISP	-0.002201916
מספר משחקים App	-0.1325883

המשתנים אותם אנו שוקלים להסיר מהמודל הינם: ממוצע בישולים למשחק AAPG, אחוז מסירות מוצלחות למשחק PS ואיבודי כדור למשחק DISP.

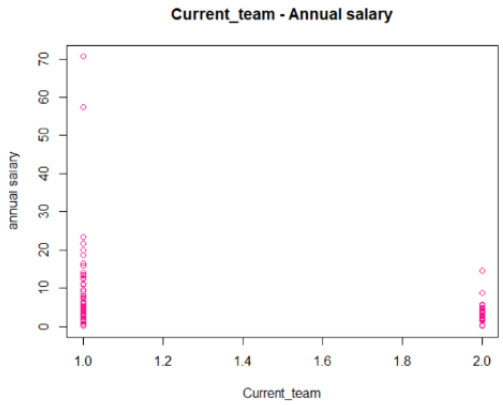
בנוסף לבדיקה בין המשתנה המסביר למשתנה המוסבר, ביצענו בדיקה האם קיימת התאמה בין המשתנים המסבירים לבין עצמם. זאת על מנת לבדוק האם קיים מתאם בין צמד של משתנים מסבירים במודל שכך הוספת משתנה מסביר לא תעניק לי אינפורמציה נוספת על מנת לאמוד את המשנה המוסבר. את בדיקה זו ביצענו גם כן באמצעות מטריצת הקורלציות של מקדם מתאם פירסון (נספח 3.2.2). לא הבחנו במקדם מתאם גבוה בין זוג משתנים מסבירים שדורש הסרת משתנים נוספת. לכן בשלב זה לא הוסרו משתנים נוספים.

3.3 התאמת משתנים

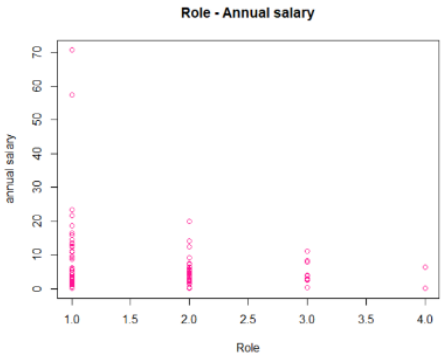
נבחרת- הנתונים שנלקחו לצורך בניית המודל דגמו שחקנים מ-34 נבחרות שונות. בשל העובדה שקיימות הרבה נבחרות שונות בטבלת הנתונים, החלטנו לחלק את הנבחרות ל-4 קטגוריות לפי היבשת אליה הנבחרת שייכת: אירופה, אמריקה, אפריקה ואסיה. אנו סבורים כי חלוקה זו תפשט את העבודה עם משתנה זה במודל.

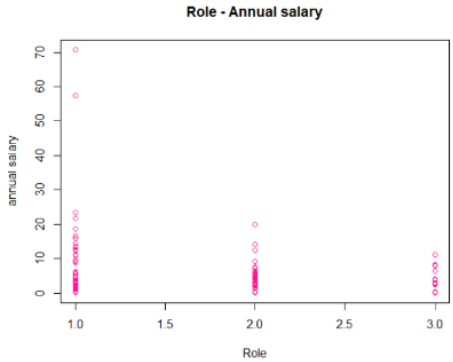


קבוצה – לצורך בניית המודל, נדגמו 100 שחקנים המשחקים ב-41 קבוצות שונות. את הקבוצות הללו בחרנו לחלק ל-2 קטגוריות עיקריות על סמך התקציב של כל קבוצה. כמו כן, נגדיר את שתי הקטגוריות הללו כ- קבוצות מדרג בכיר וקבוצות מדרג משני בהתאמה. קטגוריה 1 מכילה את כלל קבוצות הכדורגל בעלות תקציב גבוה, בעוד שקטגוריה 2 מכילה את כלל הקבוצות אשר התקציב שלהן נמוך יחסית. לפי תרשים הקטגוריות מול השכר השנתי, ניתן לראות כי נתוני השכר השנתי של השחקנים בקבוצות המשתייכות לקטגוריה 1 דומים, וזהו המצב גם עבור השחקנים בקבוצות קטגוריה 2. בנוסף, תרשים השכר השנתי כפונקציה של הקטגוריות הראה כי השכר של שחקנים המשחקים בקבוצות הדרג הבכיר גבוה יותר ביחס לשכר השחקנים המשחקים בקבוצות הדרג המשני. לכן, מצאנו לנכון לאחד קטגוריות אלו, על מנת להתייחס לקבוצות דומות כאל מקשה אחת ועל מנת לעבוד עם מספר נמוך יחסית של משתנים במודל שלנו.



תפקיד- תפקיד הינו משתנה קטגוריאלי המחולק ל-4 קטגוריות (התקפה-1, קישור-2, הגנה-3, שוער-4). על פי תרשים ה-Role אל מול ה-Annual Salary ניתן לראות את דמיון הנתונים בין תפקיד ההגנה לבין תפקיד השוער (3,4) . לכן, נבחר לאחד פעם נוספת את השחקנים הרלוונטיים לקטגוריית הגנה ונסמנה בספרה 3.





מספר משחקים- החלטנו לשנות את המשתנה מקטגוריאלי בדיד למשתנה רציף. זהו משתנה כמותי שעתיד להשתנות אצל כל שחקן ושחקן בהתאם לכושר הגופני שלו, המצב הבריאותי שלו (האם פצוע/כשיר לשחק), החלטת מאמנו על הכללתו בסגל הקבוצה וכדומה. משתנה זה עשוי לתרום לאומדן ערך השכר השנתי שלו.

במסגרת הגדרת המודל, מלבד השינויים אותם ביצענו לא ראינו לנכון להגדיר מחדש משתנים נוספים. אופן איחוד המשתנים נעשה באופן שישקף את המציאות ויצמצם את מספר המשתנים איתם נרצה לעבוד במודל, כפי שפורט עבור כל משתנה.

3.4 הגדרה של משתני דמה

על מנת לתת ייצוג של המשתנים הקטגוריאליים למודל הרגרסיה כמשתנים מסבירים נמיר אותם למשתנה דמה . בחרנו לבצע זאת על ידי ה'גישה הסטנדרטית' כפי שנלמד בכיתה. בשיטה זו נבחר את אחת הרמות כקבוצת ייחוס ולכל שאר הרמות נגדיר משנה דמה בינארי.

משתנה מסביר קבוצה:

לאחר התאמת המשתנים משתנה "קבוצה" (Current Team) קיימות שתי קטגוריות :

1- קבוצה בדרג בכירה 2- קבוצה בדרג נמוך

נגדיר את (1) קבוצה בדרג בכיר כקבוצת ייחוס ונוסיף משנה בינארי אחד:

$$CT_{2,i} = \begin{cases} 1 & \text{שחקן } i \text{ נמצא בדרג נמוך} \\ 0 & \text{אחרת} \end{cases}$$

משתנה מסביר נבחרת:

לאחר התאמת המשתנים משתנה "נבחרת" (National Team) קיימות ארבע קטגוריות על פי יבשות: 1- אירופה 2- אמריקה 3- אפריקה 4-אסיה

נגדיר את (1) אירופה כקבוצת ייחוס ונוסיף שלושה משנים בינאריים:

$$NT_{2,i} = \begin{cases} 1 & \text{שחקן } i \text{ נמצא בנבחרת ביבשת אמריקה} \\ 0 & \text{אחרת} \end{cases}$$

$$NT_{3,i} = \begin{cases} 1 & \text{שחקן } i \text{ נמצא בנבחרת ביבשת אפריקה} \\ 0 & \text{אחרת} \end{cases}$$

$$NT_{4,i} = \begin{cases} 1 & \text{שחקן } i \text{ נמצא בנבחרת ביבשת אסיה} \\ 0 & \text{אחרת} \end{cases}$$

משתנה מסביר תפקיד:

לאחר התאמת המשתנים משתנה "תפקיד" (Role) קיימות שלוש קטגוריות על פי סוג התפקיד במגרש: 1- תפקיד התקפה 2-תפקיד קישור 3- תפקיד הגנה

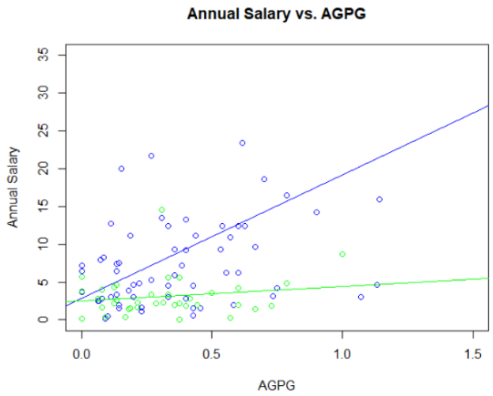
נגדיר את (1) תפקיד התקפה כקבוצת ייחוס ונוסיף שני משנים בינאריים:

$$R_{2,i} = \begin{cases} 1 & \text{שחקן } i \text{ משחק בתפקיד קישור} \\ 0 & \text{אחרת} \end{cases}$$

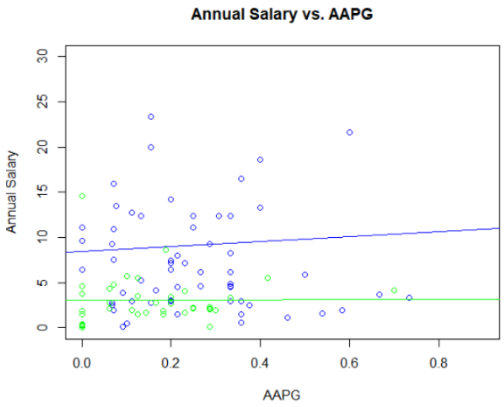
$$R_{3,i} = \begin{cases} 1 & \text{שחקן } i \text{ משחק בתפקיד הגנה} \\ 0 & \text{אחרת} \end{cases}$$

3.5 הוספת משתני אינטראקציה עבור משתני דמה

כעת, לאחר הגדרת משתני הדמה, נרצה להוסיף משתני אינטראקציה מתאימים למודל. משתני אינטראקציה הינם משתנים המביעים את התרומה השולית על השיפוע במודל. לכן, נרצה לבחון בהתחשב בקטגוריות שהגדרנו בעבודה, אילו קטגוריות משפיעות על השיפוע של קבוצת הבסיס. נרצה להוסיף עבורם משתנה אינטראקציה. נבצע בדיקה עבור כל משתני הדמה מול כל המשתנים המסבירים על ידי תרשים פיזור , כאשר הקו הכחול מייצג את השחקנים המשחקים בקבוצות מדרג בכיר יותר, בעוד שהקו הירוק מייצג את השחקנים מהקבוצות בדרג נמוך יותר. נבחן האם קיים קשר בינו לבין מידת ההשפעה של כל משתנה מסביר רציף על המשתנה המוסבר, וכך נראה את התרומה השולית של משתנה הדמה לשיפוע.

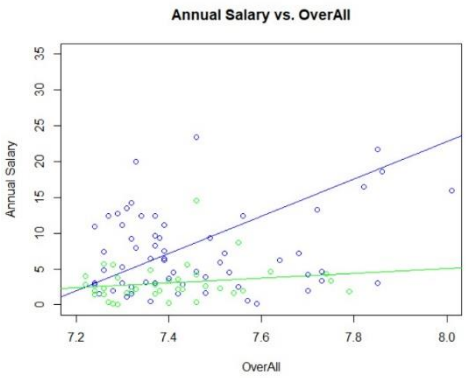


א. **<u>מספר גולים ממוצע למשחק-</u>** משתנה אשר משפיע על ערך משכורת השחקן (המשתנה המוסבר), בצורה שונה עבור דירוג הקבוצה שבה הוא משחק. ניתן לראות בבירור את המגמה כי שחקנים המשחקים בקבוצות מדרג בכיר וממוצע השערים גבוה ירוויחו יותר מאשר שחקנים עם אותו ממוצע שערים אך מקבוצות בדרג נמוך יותר. בנוסף, שתי הקטגוריות הן בעלות שיפוע שונה ולכן נרצה להוסיף משתנה אינטראקציה אשר יבטא את ההשפעה השולית של הקבוצה השנייה על שיפוע קבוצת הבסיס.

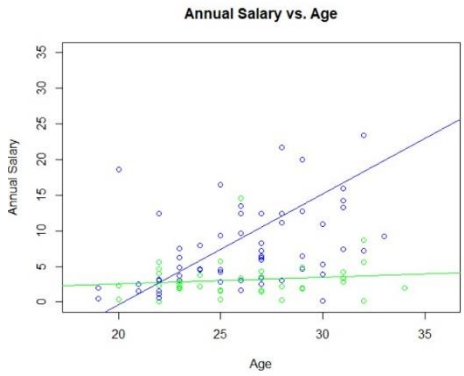


ב. **<u>מספר בישולים ממוצע למשחק -</u>** משתנה אשר משפיע על ערך השחקן (המשתנה המוסבר), אך בצורה שונה עבור דירוג הקבוצה שבה הוא משחק. ניתן לראות כי עבור ממוצע בישולים 0-0.3 ערך משכורת השחקן המשתייך לקבוצה מדרג 2 לרוב נמוכה יותר מאשר ערך השחקן בקבוצות מדרג 1. קו המגמה הלינארי יחסית מתון- דבר שמעיד על קשר פחות חזק בין ממוצע הבישולים למשחק אל מול משכורות השחקנים. בנוסף, שתי הקטגוריות הן בעלות שיפוע שונה ולכן נרצה להוסיף משתנה אינטראקציה אשר יבטא את ההשפעה השולית של הקבוצה השנייה על שיפוע קבוצת הבסיס.

ג. **<u>הדירוג הכללי של השחקן-</u>** משתנה המשפיע על ערך השחקן בצורה שונה בהינתן דירוג הקבוצה שבה הוא משחק. ניתן לראות כי שיפוע הקו כחול תלול יותר וגבוה יותר מאשר של הקו הירוק, ואכן בגרף רוב נקודות השכר השנתי גבוהות יותר כאשר השחקן משתייך למועדון מדרג 1. מכיוון ששתי הקטגוריות הן בעלות שיפוע שונה ולכן נרצה להוסיף משתנה אינטראקציה אשר יבטא את ההשפעה השולית של הקבוצה השנייה על שיפוע קבוצת הבסיס.



ד. **<u>גיל השחקן-</u>** משפיע על ערך השחקן בהתאם לדירוג הקבוצה שבה הוא משחק. לפי הגרף, שיפוע הקו כחול תלול יותר מאשר של הקו הירוק, ולרוב נמצא מעליו. בהתאם לכך, רוב נקודות השכר השנתי גבוהות יותר כאשר השחקן משתייך למועדון מדרג 1 ככל שהשחקן עולה בגילו. מכיוון ששתי הקטגוריות הן בעלות שיפוע שונה ולכן נרצה להוסיף משתנה אינטראקציה אשר יבטא את ההשפעה השולית של הקבוצה השנייה על שיפוע קבוצת הבסיס.



<u>משתנה האינטראקציה:</u>

$$CT_2 * X_2 = \begin{cases} X_2, & if\ the\ Team\ is\ 2 \\ 0\ , & else \end{cases}$$

לאחר כלל העיבוד המקדים שביצענו, מודל הרגרסיה הראשוני שהתקבל הינו:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}Role_2 + \widehat{\beta_3}Role_3 + \widehat{\beta_4}CT_2 + \widehat{\beta_5}NT_2 + \widehat{\beta_6}NT_3 + \widehat{\beta_7}NT_4 + \widehat{\beta_8}X_2 + \widehat{\beta_9}X_3 + \widehat{\beta_{10}}X_4 + \widehat{\beta_{11}}X_5 + \widehat{\beta_{12}}X_6 + \widehat{\beta_{13}}X_7 + \widehat{\beta_{14}}CT_2 * X_1 + \widehat{\beta_{15}}CT_2 * X_2$$

כאשר:

X_1 – שערים של שחקן בממוצע למשחק

X_2 – בישולים של שחקן בממוצע למשחק

X_3 –אחוז מסירות מוצלחות של שחקן למשחק

X_4 –איבודי כדור של השחקן

X_5 – דירוג כללי של השחקן

X_6- גיל השחקן

X_7 – מספר משחקים של שחקן

4. התאמת המודל ובדיקת הנחות המודל

4.1 בחירת משתני המודל

נשתמש באלגוריתמים השונים שלמדנו בכיתה בכדי לבחור את המשתנים אותם נכלול במודל הסופי שלנו. נבחן את המודלים לפי הקריטריונים הבאים: R^2 , R^2_{adj} , AIC ו – BIC. זאת ע"פ הגישות הבאות: Forward Selection, Backward Elimination, Stepwise Regression.

- **מדד** R^2_{adj} -מדד זה מייצג את אחוז השונות המוסברת במודל. זה מותאם למודל רגרסיה מרובה בכך שמתייחס למספר המשתנים המוסברים (דרגות החופש). נרצה לבחור את המודל בעל הערך המקסימאלי כלומר אחוז השונות המוסברת הוא הגדול ביותר (שואף לערך 1).
- **מדדי AIC, BIC –** מדדים אלו בוחנים את טיב ההתאמה של המודל לנתונים מבחינת הנראות שלו (ככל שהנראות גדלה, דבר זה יגרור ערך מדדים קטן), כמו כן דבר זה "קונס" את המודל על פי מספר הפרמטרים שבו - במדד AIC, ועל פי מספר התצפיות - במדד BIC. לכן, נרצה למצוא את הערכים המינימליים עבור המודל המתאים ביותר.

המודל המלא שקיבלנו לפני הסרת המשתנים וטרם הרצת האלגוריתמים:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x5 * DataBase$CurrentTeam.cat + x6 * DataBase$CurrentTeam.cat)

Residuals:
    Min      1Q  Median      3Q     Max
-16.032  -2.874   0.089   2.383  34.417

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -167.2751    34.1337  -4.901 4.39e-06 ***
x1                              4.8357     3.7970   1.274 0.206215
x2                             -3.8446     5.1693  -0.744 0.459041
x3                              0.1291     0.1251   1.032 0.304866
x4                             -1.1184     1.3714  -0.816 0.416999
x5                             19.4335     4.8623   3.997 0.000134 ***
x6                              1.1562     0.2688   4.301 4.43e-05 ***
x7                             -0.4696     0.3595  -1.306 0.194872
x8                             -1.6548     1.3726  -1.206 0.231246
x9                             -0.3779     0.9647  -0.392 0.696198
DataBase$CurrentTeam.cat2     173.4662    67.4387   2.572 0.011802 *
x5:DataBase$CurrentTeam.cat2  -19.5935     9.0463  -2.166 0.033053 *
x6:DataBase$CurrentTeam.cat2   -1.2166     0.4472  -2.721 0.007869 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.113 on 87 degrees of freedom
Multiple R-squared:  0.5269,	Adjusted R-squared:  0.4616
F-statistic: 8.073 on 12 and 87 DF,  p-value: 5.928e-10
```

Forward Selection- האלגוריתם מתחיל במודל הריק ומכניס אליו משתנה אחד בכל איטרציה. המשתנה המובהק ביותר (בעל ה-F_{st} החלקי הגדול ביותר) ייכנס למודל. נעצור כאשר לא נדחה את השערת האפס עבור המועמד הטוב ביותר להיכנס למודל.

המודל הנבחר עבור אלגוריתם זה הינו:

$$\hat{y} = -187.6826 + 22.0329x_5 + 1.2282x_6 + 163.4543 * CT_2 - 18.6896x_5 * CT_2 - 1.1302x_6 * CT_2$$

Backward Elimination- במודל זה אנו מתחילים עם המודל המלא המכיל את כלל המשתנים. האלגוריתם יוציא בכל איטרציה את המשתנה הכי פחות מובהק (בעל ה- F_{st} החלקי הקטן ביותר). נעצור כאשר לא נדחה את השערת האפס עבור המועמד הטוב ביותר לצאת מהמודל.

המודל הנבחר עבור אלגוריתם זה הינו:

$$\hat{y} = -187.6826 + 22.0329x_5 + 1.2282x_6 + 163.4543 * CT_2 - 18.6896x_5 * CT_2 - 1.1302x_6 * CT_2$$

Stepwise Regression- נשלב את השיטות אשר ביצענו מעלה. בכל שלב נבדוק האם להכניס או להוציא משתנים שנוספו למודל בצעדים הקודמים. כתוצאה מכך משתנה מסביר שנוסף בצעדים הקודמים יכול להפוך גם למיותר בגלל הקשרים עם משתנים המסבירים האחרים אשר נוספו בינתיים למודל.

בשלושת השיטות התקבל הפלט הבא:

```
Call:
lm(formula = y ~ x5 + x6 + DataBase$CurrentTeam.cat + x5:DataBase$CurrentTeam.cat +
    x6:DataBase$CurrentTeam.cat)

Residuals:
    Min      1Q  Median      3Q     Max
-16.285  -2.473  -0.375   2.599  37.195

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -187.6826    27.4265  -6.843 7.82e-10 ***
x5                             22.0329     3.7585   5.862 6.74e-08 ***
x6                              1.2282     0.2528   4.859 4.71e-06 ***
DataBase$CurrentTeam.cat2     163.4543    65.2558   2.505  0.01397 *
x5:DataBase$CurrentTeam.cat2  -18.6896     8.7395  -2.139  0.03507 *
x6:DataBase$CurrentTeam.cat2   -1.1302     0.4064  -2.781  0.00654 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.081 on 94 degrees of freedom
Multiple R-squared:  0.4933,    Adjusted R-squared:  0.4664
F-statistic: 18.31 on 5 and 94 DF,  p-value: 1.193e-12
```

המודל הנבחר עבור אלגוריתם זה הינו :

$$\hat{y} = -187.6826 + 22.0329x_5 + 1.2282x_6 + 163.4543 * CT_2 - 18.6896x_5 * CT_2 - 1.1302x_6 * CT_2$$

נבחן את שלושת המודלים שהתקבלו באמצעות המדדים הבאים:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}, R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$AIC = n\log(SSE/n) + 2(k+1)$$

$$BIC = n\log(SSE/n) + \log(n)(k+1)$$

להלן טבלה המסכמת את נתוני המדדים: (פלט וקוד בנספח 4.1.1)

Stepwise	Backward	Forward	Full	
0.4664	0.4664	0.4664	0.4616	$\mathbf{R^2_{adj}}$
683.0906	683.0906	683.0906	690.2475	**AIC**
701.3268	701.3268	701.3268	726.7199	**BIC**

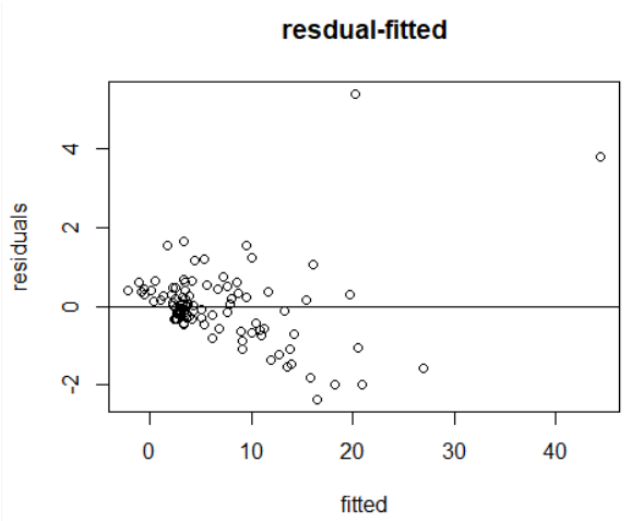
ניתן לראות כי התקבל אותו מודל באמצעות שלושת השיטות. לכן, המודל הסופי הינו:

$$\hat{y} = -187.6826 + 22.0329x_5 + 1.2282x_6 + 163.4543 * CT_2 - 18.6896x_5 * CT_2 - 1.1302x_6 * CT_2$$

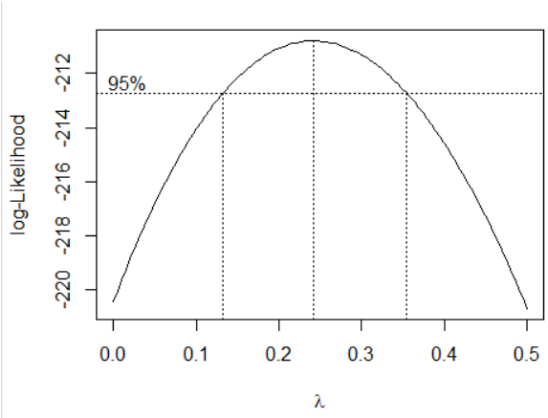
4.2 בדיקת הנחות המודל

נבדוק האם המודל מקיים את 3 ההנחות: ליניאריות, שוויון שונויות ונורמליות השגיאות.

<u>בדיקת הנחת שוויון שוניות:</u>
נפיק תרשים פיזור הנחת שוויון השונויות (נספח קוד 4.2.1). תרשים שאריות בודק לינאריות ושוויון שונויות. נבדוק בתרשים השאריות האם השאריות שאנו רואים הן סימטריות ביחס לקו ה – 0 וכמו כן שקיים פיזור אחיד יחסית. כלומר, נסתכל עבור כל ערך חזוי האם קיימת אותה השונות ובאזור זה נצטרך לראות פיזור בקירוב אחיד סביב ה - 0.

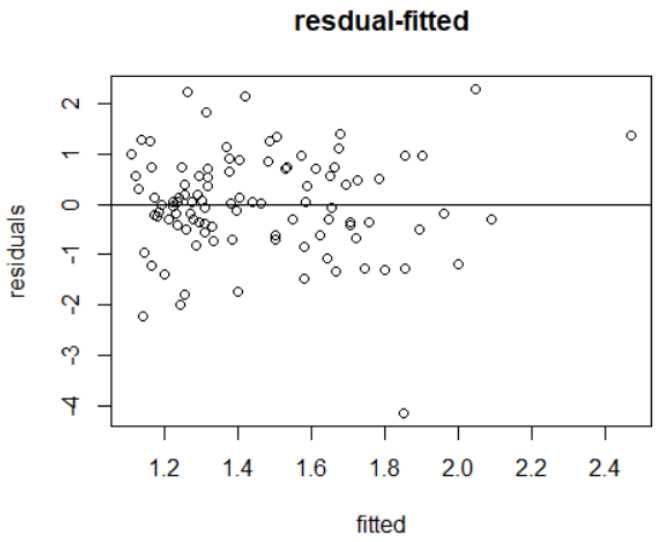


ניתן לראות כי עבור המודל הנוכחי, פיזור הנקודות בגרף סביב האפס אינו אחיד. לכן, ניתן להסיק שהנחת שוויון השונויות לא מתקיימת. נרצה לבצע טרנספורמציה על Y על מנת לגרום למודל לקיים הנחה זו.



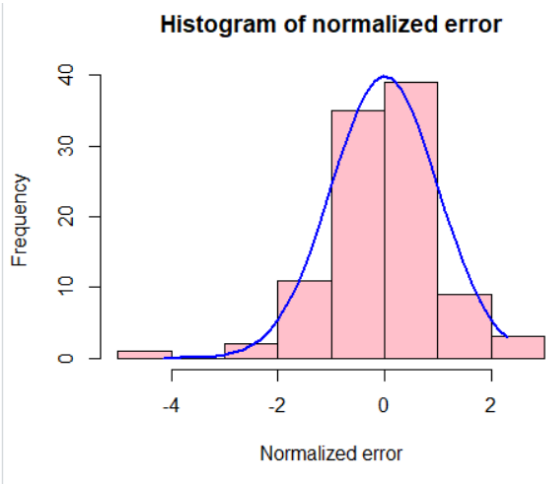
באמצעות תרשים COX-BOX ניתן לראות כי הלמבדה ממקסמת את פונקציית הנראות עבור המודל שלנו היא בערך 0.25.

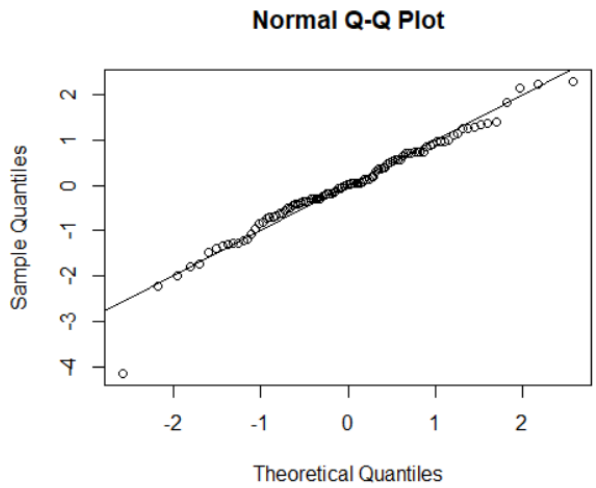
ניקח את הלמבדה הנ"ל ונעלה את המשתנה המוסבר שלנו (Y) בחזקת הערך שלו על מנת לקבל מודל המקיים שוויון שונויות. נבחן את המודל החדש באמצעות תרשים פיזור השגיאות כדי לוודא קיום הנחה זו:



מהתרשים ניתן לראות כי ישנו פיזור די שווה ובעל שונות גדולה לאורך ציר X (y̓).
כלומר, פיזור התצפיות סביב קו התוחלת של השגיאות הינו אחיד לאורך ציר ŷ ולכן ניתן להסיק שהנחת שוויון השונויות אכן מתקיימת.

<u>הנחת הנורמליות:</u>





על מנת להסיק שהנחת הנורמליות מתקיימת, נרצה לראות שב-היסטוגרמה שלנו יתקבל פעמון שמדמה התפלגות נורמלית. בתרשים ה - QQ-plot נרצה לקבל תצפיות אשר נמצאות על קו ה – 45 מעלות הלינארי.

מההיסטוגרמה אנו אכן רואים התנהגות אשר דומה לפעמון ולכן נשער כי במודל שלנו הנחת הלינאריות מתקיימת. בנוסף, בתרשים הכמותונים ניתן לראות שהתצפיות נמצאות באופן יחסי על הקו הלינארי (למעט תצפיות בודדות בקצוות הערכים

הקיצוניים ומעט במרכז). לכן, עדיין לא נסיק שהנחת הנורמליות מתקיימת, אך על מנת לוודא את התקיימותה נבצע מבחן קולמוגורוב סמירנוף (נספח 7.4):

```
        One-sample Kolmogorov-Smirnov test

data:  DataBase$stan_residuals
D = 0.051069, p-value = 0.9567
alternative hypothesis: two-sided
```

במבחן זה השערת האפס הינה שהנתונים אכן מתפלגים נורמלית. במבחן זה כאמור קיבלנו ש – P-value גדול מאוד מר"מ שקבענו 0.05 ולכן לא נדחה את השערת האפס. כלומר, נסיק שהשגיאות המתוקננות מתפלגות נורמלית.

<u>הנחת הלינאריות:</u>

כדי לבחון את הנחת הלינאריות נתייחס לתרשים פיזור השאריות (אשר הוצג למעלה). בתרשים זה נצפה לראות פיזור אקראי ביחס לציר ה – 0 ובנוסף נרצה לראות כי לא קיימת מגמה. מתרשים פיזור השאריות שלנו ניתן לראות שהשאריות מפוזרות סביב קו האפס באופן סימטרי יחסית. כמו כן, לא קיימת סוג של מגמה. לכן, נסיק שהנחת הלינאריות מתקיימת.

4.3 דוגמא לשימוש במודל הנבחר

בעלי הקבוצה מכבי חיפה יענקלה שחר מעוניין לרכוש שחקן חדש לקבוצתו. הוא מתעניין ברכישתו של השחקן מאור בוזגלו. השחקן אינו מופיע במאגר הנתונים על בסיסו בנינו את המודל. יענקלה שחר צריך עזרה בלאמוד את משכורתו השנתית של מאור ולפי ערך זה להחליט האם שווה לקנות את השחקן. נשתמש במודל הרגרסיה שבנינו ונבדוק מה הערך שהתקבל:

$$\hat{y} = -187.6826 + 22.0329x_5 + 1.2282x_6 + 163.4543 * CT_2 - 18.6896x_5 * CT_2 - 1.1302x_6 * CT_2$$

Name	App	CT	NT	Role	Age	AGPG	AAPG	PS	DISP	OA
Maor Buzaglo	15	1	1	1	33	0.40	0.07	69	0.9	7.32

להלן ערך השחקן ע"פ המודל שלנו:

$$\hat{y} = -187.6826 + 22.0329 * 7.32 + 1.2282 * 33 + 163.4543 * 1 - 18.6896 * 7.92 * 1 - 1.1302 * 33 * 1 = 3.41$$

קיבלנו כי משכורתו של השחקן מוערכת כיום בכ- 3.41 מיליון יורו לשנה.

4.4 ביצוע השערה המבוססת על הנחות המודל

רשמי: אינטר החתימה את אשלי יאנג מיונייטד

נבחר לבצע מבחן השערות עבור תצפית בודדת עתידית. השחקן האנגלי אשלי יאנג שמשחק בקבוצת מנצ'סטר יונייטד האנגלית עבר לקבוצת הכדורגל האיטלקית אינטר מילאנו (על פי דיווחי אתר ספורט 5 מצורף משמאל תמונה מתוך העמוד הראשי) . נרצה לבדוק על פי מודל הרגרסיה שבנינו האם המעבר היא כדאי לו מבחינה כלכלית, כלומר האם השכר השנתי שלו גדל בחוזה החדש עליו חתם באינטר.
נרצה לאמוד את השכר השנתי אם כן ולוודא האם היה גבוה יותר משכרו במנצ'סטר יונייטד שעמד על 6.24 מיליון יורו.

<u>נתוני המודל על השחקן: (נלקחו מאותם אתרים מהם לקחנו את הנתונים בחלק א'):</u>

$X_5 = 6.92$ – דירוג כללי של השחקן

$X_6 = 34$ -גיל השחקן

$CT_2 = 0$ -הקבוצה הנוכחית של השחקן

נבצע מבחן השערות ברמת מובהקות 5%. ההשערות יהיו:

$$H_0: y_0 \geq 6.24$$

$$H_1: else$$

אומדן השכר השנתי לפי המודל הינו:

$$\hat{y}_0 = -187.6826 + 22.0329 * 6.92 + 1.2282 * 34 + 163.4543 * 0 - 18.6896 * 6.92 * 0 - 1.1302 * 34 * 0 = 6.543$$

נבנה את סטטיסטי המבחן על פי הנוסחה:

$$t_{s.t} = \frac{y_0 - \hat{y}_0}{\sqrt{v(\widehat{y} - y_0)}} = \frac{y_0 - \hat{y}_0}{\sqrt{\hat{\sigma}^2(1 + c * (X^T X)^{-1} * c^T)}} = \frac{6.24 - 6.543}{\sqrt{7.081^2(1 + c * (X^T X)^{-1} * c^T)}}$$

לשם חישוב ה-t הסטטיסטי נשתמש במדדים: (נספח 4.4.1)

1. סטיית התקן של השגיאות בריבוע (על סמך פלט R):

$$\hat{\sigma}^2 = 7.081^2 = 50.14$$

2. הגדרת את וקטור C המכיל את נתוני המסבירים:

$$c^T = [1, 6.92, 34, 0]$$

נחשב את הערך הסטטיסטי

$$t_{st} = -0.0115$$

נמצא ערך קריטי מטבלת t :

$$t_{cr} = t_{n-k-1,1-\alpha} = t_{96,0.95} = 1.66$$

נקבל כי: $\boldsymbol{tst < tcr}$

נראה כי t_{st} לא באזור הדחייה, ולכן נאמר בר"מ של 5% כי אין מספיק נתונים לדחות את השערת האפס. כלומר, לא נוכל לקבוע האם יותר שווה לו לעבור לאינטר מבחינת שכר שנתי על סמך המודל הנתון.

5. שיפור המודל

כעת, נבדוק את מתאמי פירסון עבור המשתנים המסבירים שהוסרו מהמודל (אשר היו בעלי מתאם נמוך) ועבור המשתנים המסבירים שנשארו שאותם נשאף לשפר עוד יותר. ננסה להשיג שיפור במדד עבור משתנים אלו באמצעות טרנספורמציה עליהם. בחנו את המתאמים של כל המסבירים (נספח 5.1) וניתן לראות כי מתאם פירסון של המשתנים שלא כללנו במודל הסופי (מלבד AGPG), נשאר נמוך יחסית גם לאחר הטרנספורמציה ולכן נשאיר אותם מחוץ למודל. עם זאת, קיים שיפור במתאם עבור שלושה משתנים על ידי טרנספורמציה.

להלן פרטי הטרנספורמציה על המשתנים אשר הניבו שיפורים במתאם פירסון:

משתנה	טרנספורמציה נבחרת	מקדם פירסון לפני הטרנספורמציה	מקדם פירסון לאחר טרנספורמציה
X1 – AGPG	ריבועית	0.3802837	0.4296012
X5 – OverAll	אקספוננציאלית	0.5289814	0.6354754
X6 – Age	אקספוננציאלית	0.3859991	0.4054136

אנו רואים כי רק עבור המשתנה X1 (ממוצע השערים למשחק) מקדם פירסון עלה ועל כן החלטנו להוסיפו לבדיקת המודל. כמו כן, הרצנו בדיקות רגרסיה לפנים, לאחור ובצעדים על המודל המלא (שכולל את כל המשתנים) בתוספת המשתנים לאחר הטרנספורמציה, כדי לקבל מודל אידיאלי. להלן המודל שהתקבל:

```
Call:
lm(formula = y ~ exp(x5) + exp(x6) + DataBase$CurrentTeam.cat +
    exp(x5):DataBase$CurrentTeam.cat + exp(x6):DataBase$CurrentTeam.cat)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1687  -2.3750  -0.6811   2.2684  15.2391

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -1.110e+01  1.890e+00  -5.873 6.44e-08 ***
exp(x5)                           1.019e-02  9.730e-04  10.475  < 2e-16 ***
exp(x6)                           7.983e-14  8.331e-15   9.583 1.42e-15 ***
DataBase$CurrentTeam.cat2         1.140e+01  5.725e+00   1.991   0.0494 *
exp(x5):DataBase$CurrentTeam.cat2 -8.510e-03  3.347e-03  -2.542   0.0126 *
exp(x6):DataBase$CurrentTeam.cat2 -8.026e-14  1.215e-14  -6.606 2.35e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.143 on 94 degrees of freedom
Multiple R-squared:  0.7327,    Adjusted R-squared:  0.7185
F-statistic: 51.53 on 5 and 94 DF,  p-value: < 2.2e-16
```

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \exp(x_5) + \hat{\beta}_2 \exp(x_6) + \hat{\beta}_3 Team2 + \hat{\beta}_4 \exp(x_5) * Team2 + \hat{\beta}_5 \exp(x_6) * Team2$$

ניתן לראות כי מדד R^2_{adj} עבור המודל המתואר השתפר לפי כל הבדיקות, עד לכדי 71.85%.

שיטה	R^2_{adj}
לפני טרנספורמציה	0.594
Forward	0.7185
Backward	0.7185
Stepwise	0.7185

6. מסקנות והמלצות

בעבודה זו רצינו לבצע מידול לשכרו של שחקן כדורגל על ידי כלי הרגרסיה הלינארית. בחרנו עבור המודל 10 משתנים מסבירים שונים אשר יכולים להשפיע על שכרו של השחקן. לאחר הרצת הרגרסיה כפי שנלמד בכיתה הגענו למודל המושפע משלושה פרמטרים מרכזיים אצל השחקן והקשר בניהם (הקבוצה הנוכחית בה משחק, דירוג הכללי של השחקן וגילו) ועל ידי כך התקבל המודל הבא:

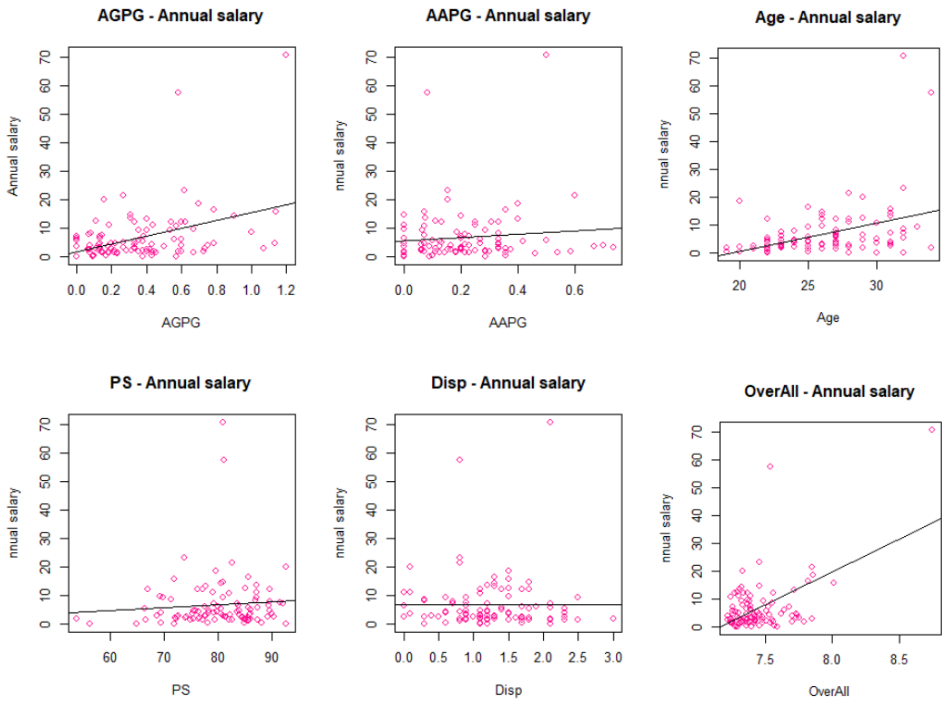
$$\hat{y} = -187.6826 + 22.0329x_5 + 1.2282x_6 + 163.4543CT$$
$$- 18.6896x_5 * CT - 1.1302x_6 * CT$$

המצלות להמשך:

במידה ונרצה לבצע מודל מדויק יותר עבור השכר של שחקן נמליץ לבצע איסוף נתונים גדול יותר ולקבל את המידע על מספר רב של שחקנים. כמו כן במסגרת איסוף נתונים אלה נמליץ להרחיב את המגוון ממנו נלקחו הנתונים. בעבודה זו נלקחו נתונים מהליגות הבכירות ולכן יתכן כי קיימת הטיה של המקדמים בהתאמה למידע זה ולכן גדלת המגוון כמו מדינות נוספות שאינן נמצאות בבסיס הנתונים וכן ליגות נמוכות יותר יכול לגרום הגברת הדיוק של המודל. כמו כן נמליץ להוסיף משנים מסבירים אשר לא היו חשופים לנו אך יכולים להימצא במחקר רחב יותר. למשל, ההכנסה שנתית ממכירת מוצרי שחקן בחנויות המועדון לו הוא שייך משום שפרמטר זה הינו משפיע על המועדון לקנות שחקן בכוונה שיגדיל את כמות המכירות הנ"ל.

7. נספחים

3.2.1 תרשים מתאם בין המשתנה המוסבר לכל אחד מהמסבירים שאינו קטגוריאלי




```
par(mfrow=c(2,2))

plot(x=DataBase$AGPG,y=DataBase$`Annual salary`,col = "deeppink", main =
'AGPG - Annual salary', xlab='AGPG', ylab='Annual salary')

linearmodel2<-lm(Annual_salary ~ AGPG, data = DataBase)

abline(linearmodel2)


plot(x=DataBase$AAPG,y=DataBase$`Annual salary`,col= "deeppink", main = 'AAPG
- Annual salary', xlab='AAPG', ylab='nnual salary')

linearmodel2<-lm(Annual_salary  ~ AAPG, data = DataBase)

abline(linearmodel2)
```

מודל רגרסיה מרובה

```
Call:
lm(formula = Annual_salary ~ AGPG + AAPG + PS + Disp + OverAll +
    Age + App, data = DataBase)

Residuals:
    Min      1Q  Median      3Q     Max
-14.321  -4.195  -0.159   2.666  39.426

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -158.0933    33.1269  -4.772 6.83e-06 ***
AGPG           7.3036     3.7366   1.955  0.05367 .
AAPG          -4.5347     5.0083  -0.905  0.36760
PS             0.1736     0.1222   1.420  0.15894
Disp          -1.3657     1.3512  -1.011  0.31479
OverAll       18.5803     4.5042   4.125 8.12e-05 ***
Age            0.7343     0.2170   3.384  0.00105 **
App           -0.4897     0.3817  -1.283  0.20264
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.641 on 92 degrees of freedom
Multiple R-squared:  0.4226,    Adjusted R-squared:  0.3787
F-statistic: 9.621 on 7 and 92 DF,  p-value: 6.297e-09
```

3.2.2 מטריצת הקורלציות של מקדם מתאם פירסון

	App	Current team	National team	Role	Age	AGPG	AAPG	PS	Disp	OverAll	Annual salary
App	1.000000000	0.167966820	-0.065778897	-0.036232790	-0.004993781	0.03946828	-0.15995493	-0.219529953	-0.121145176	-0.061800674	-0.132588285
Current team	0.167966820	1.000000000	0.042043116	-0.006107732	-0.002507855	-0.11904847	-0.23956834	-0.290049809	0.206244932	-0.161663868	-0.303511904
National team	-0.065778897	0.042043116	1.000000000	-0.040896291	0.056760261	-0.06387784	0.01494798	0.167661997	0.134152076	-0.092799393	-0.005622231
Role	-0.036232790	-0.006107732	-0.040896291	1.000000000	-0.195564064	-0.50209677	-0.21023965	0.407565487	-0.414860823	-0.251638133	-0.198744478
Age	-0.004993781	-0.002507855	0.056760261	-0.195564064	1.000000000	0.21603114	-0.02046701	-0.056878431	-0.014874011	0.150752726	0.380283693
AGPG	0.039468283	-0.119048473	-0.063877840	-0.502096772	0.216031142	1.00000000	0.10872519	-0.401741495	0.318703917	0.511156631	0.385999126
AAPG	-0.159954928	-0.239568343	0.014947978	-0.210239647	-0.020467006	0.10872519	1.00000000	0.078288801	0.092020408	0.349813003	0.104083516
PS	-0.219529953	-0.290049809	0.167661997	0.407565487	-0.056878431	-0.40174149	0.07828880	1.000000000	-0.326793877	-0.009488433	0.077762051
Disp	-0.121145176	0.206244932	0.134152076	-0.414860823	-0.014874011	0.31870392	0.09202041	-0.326793877	1.000000000	0.150752023	-0.002201916
OverAll	-0.061800674	-0.161663868	-0.092799393	-0.251638133	0.150752726	0.51115663	0.34981300	-0.009488433	0.150752023	1.000000000	0.528981445
Annual salary	-0.132588285	-0.303511904	-0.005622231	-0.198744478	0.380283693	0.38599913	0.10408352	0.077762051	-0.002201916	0.528981445	1.000000000

4.1.1 קוד רגרסיה לפנים, לאחור ובצעדים

```
#------------------------Forward--------------------------

lm.null<- lm(y~1,data=DataBase)
model.aic.forward<-step(lm.null,direction = "forward", trace = 1,scope = list(upper=model2),data=DataBase)
summary(model.aic.forward)
AIC(model.aic.forward)
BIC(model.bic.forward)


#------------------------Backword--------------------------

lm.null<- lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x5*DataBase$`CurrentTeam.cat`+x6*DataBase$`CurrentTeam.cat`)
model.aic.backward<-step(lm.null,direction = "backward", trace = 1,scope = list(upper=model2),data=DataBase)
summary(model.aic.backward)
AIC(model.aic.backward)
BIC(model.bic.backward)

#------------------------both--------------------------

lm.null<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x5*DataBase$`CurrentTeam.cat`+x6*DataBase$`CurrentTeam.cat`)
model.aic.both<-step(lm.null,direction = "both", trace = 1,scope = list(upper=model2),data=DataBase)
summary(model.aic.both)
AIC(model.aic.both)
BIC(model.bic.both)
```

פלט רגרסיה לפנים, לאחור ובצעדים ומדדי AIC BIC

```
Call:
lm(formula = y ~ x5 + x6 + DataBase$CurrentTeam.cat + x5:DataBase$CurrentTeam.cat +
    x6:DataBase$CurrentTeam.cat)

Residuals:
    Min      1Q  Median      3Q     Max
-16.285  -2.473  -0.375   2.599  37.195

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -187.6826    27.4265  -6.843 7.82e-10 ***
x5                             22.0329     3.7585   5.862 6.74e-08 ***
x6                              1.2282     0.2528   4.859 4.71e-06 ***
DataBase$CurrentTeam.cat2     163.4543    65.2558   2.505  0.01397 *
x5:DataBase$CurrentTeam.cat2  -18.6896     8.7395  -2.139  0.03507 *
x6:DataBase$CurrentTeam.cat2   -1.1302     0.4064  -2.781  0.00654 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.081 on 94 degrees of freedom
Multiple R-squared:  0.4933,    Adjusted R-squared:  0.4664
F-statistic: 18.31 on 5 and 94 DF,  p-value: 1.193e-12
```

4.2.1 קוד בדיקת הנחות המודל

קוד תרשים BOX-COX

```
newModle<-lm(y~x5+x6+DataBase$`CurrentTeam.cat`+x5*DataBase$`CurrentTeam.cat`+x6*DataBase$`CurrentTeam.cat`)
boxcox(newModle,lambda = seq(0, 0.5, 1/10))
improvedModel<-lm(y^(0.4)~x1+x2+x3+x4+x5+x6+x7+x8+x9+x5*DataBase$`CurrentTeam.cat`+x6*DataBase$`CurrentTeam.cat`)
summary(improvedModel)
```

קוד תרשים פיזור שגיאות של המודל המתוקן

```
newModle<-newModle<-lm(y~x5+x6+DataBase$`CurrentTeam.cat`+x5*DataBase$`CurrentTeam.cat`+x6*DataBase$`CurrentTeam.cat`
predicted<- predict(lm(newModle))
unstandardizedresiduals<- resid(newModle)
residuals<-(unstandardizedresiduals-mean(unstandardizedresiduals))/sd(unstandardizedresiduals)
plot(predicted,residuals,main="resdual-fitted", xlab="fitted" ,ylab="residuals")
abline(0,0)
```

קוד תרשים פיזור שגיאות לאחר טרנספורמציה על המשתנה המוסבר Y

```
newModle<-lm(y^(0.25)~x5+x6+DataBase$`CurrentTeam.cat`+x5*DataBase$`CurrentTeam.cat`+x6*DataBase$`CurrentTeam.cat`)
predicted<- predict(lm(newModle))
unstandardizedresiduals<- resid(newModle)
residuals<-(unstandardizedresiduals-mean(unstandardizedresiduals))/sd(unstandardizedresiduals)
plot(predicted,residuals,main="resdual-fitted", xlab="fitted" ,ylab="residuals")
abline(0,0)
```

קוד תרשים כמותונים QQPLOT לבדיקת הנחת הנורמליות

```
predicted<- predict(lm(newModle))
unstandardizedresiduals<- resid(newModle)
residuals<-(unstandardizedresiduals-mean(unstandardizedresiduals))/sd(unstandardizedresiduals)
mod<-lm(newModle)
DataBase$fitted<-fitted(mod)
DataBase$residuals<-residuals(mod)
s.e_res <- sqrt(var(DataBase$residuals))
DataBase$stan_residuals<-(residuals(mod)/s.e_res)
qqnorm(DataBase$stan_residuals)
abline(a=0, b=1)
```

קוד תרשים היסטוגרמה של השגיאות המתוקננות

```
h<-hist(DataBase$stan_residuals, breaks=8, col="pink", xlab ="Normalized error", main="Histogram of normalized error'
xfit<-seq(min(DataBase$stan_residuals),max(DataBase$stan_residuals),length=40)
yfit<-dnorm(xfit,mean=mean(DataBase$stan_residuals),sd=sd(DataBase$stan_residuals))
yfit <- yfit*diff(h$mids[1:2])*length(DataBase$stan_residuals)
lines(xfit, yfit, col="blue", lwd=2)
```

4.4.1 קוד מבחן השערות t עבור תצפית בודדת עתידית

```
#------------------------t s.t.--------------------------

X <- DataBaseNew
xt <- t(X)
xtx <- (as.matrix(xt) %*% as.matrix(X))
c <- as.matrix(c(1, 6.92, 34 , 0))
ct <- t(c)
mse <- (7.081)^2
t_st <- (6.24-6.543)/sqrt(mse*(1+(ct%*%solve(xtx)%*%c)))
t_cr<-qt(0.95,96)
paste(t_st)
paste(t_cr)
```

5.1. השוואת טרנספורמציה אפשרית על המשתנים

```
> #-----Changes in the model-----
>
> cor(Annual_salary, Age, method = c("pearson"))
[1] 0.3802837
> cor(Annual_salary, log(Age), method = c("pearson"))
[1] 0.3643413
> cor(Annual_salary, Age^2, method = c("pearson"))
[1] 0.3948324
> cor(Annual_salary, exp(Age), method = c("pearson"))########
[1] 0.4054136
>
> cor(Annual_salary, OverAll, method = c("pearson"))
[1] 0.5289814
> cor(Annual_salary, log(OverAll), method = c("pearson"))
[1] 0.5140728
> cor(Annual_salary, OverAll^2, method = c("pearson"))
[1] 0.5439114
> cor(Annual_salary, exp(OverAll), method = c("pearson"))#########
[1] 0.6354754
>
> cor(Annual_salary, AGPG, method = c("pearson"))
[1] 0.3859991
> cor(Annual_salary, log(AGPG), method = c("pearson"))
[1] NaN
> cor(Annual_salary, AGPG^2, method = c("pearson"))#########
[1] 0.4296012
> cor(Annual_salary, exp(AGPG), method = c("pearson"))
[1] 0.4187809
>
> cor(Annual_salary, AAPG, method = c("pearson"))
[1] 0.1040835
> cor(Annual_salary, log(AAPG), method = c("pearson"))
[1] NaN
> cor(Annual_salary, sqrt(AAPG), method = c("pearson"))
[1] 0.1068532
> cor(Annual_salary, exp(AAPG), method = c("pearson"))#########
[1] 0.1024646
>
> cor(Annual_salary, App, method = c("pearson"))
[1] -0.1325883
> cor(Annual_salary, log(App), method = c("pearson"))
[1] -0.1281704
> cor(Annual_salary, App^2, method = c("pearson"))
[1] -0.1355789
> cor(Annual_salary, exp(App), method = c("pearson"))#########
[1] -0.1125784
>
> cor(Annual_salary, Disp, method = c("pearson"))
[1] -0.002201916
> cor(Annual_salary, log(Disp), method = c("pearson"))
[1] NaN
> cor(Annual_salary, sqrt(Disp), method = c("pearson"))#########
[1] -0.00544202
> cor(Annual_salary, exp(Disp), method = c("pearson"))
[1] -0.0110643
```

```
>
> cor(Annual_salary, PS, method = c("pearson"))
[1] 0.07776205
> cor(Annual_salary, log(PS), method = c("pearson"))
[1] 0.08541076
> cor(Annual_salary, PS^2, method = c("pearson"))
[1] 0.06991245
> cor(Annual_salary, exp(PS), method = c("pearson"))########
[1] 0.03426836
```