

פרויקט ברגרסיה לינארית - חלק א'

גיל האישה בעת הלידה הראשונה והגורמים שמשפיעים על כך



קבוצה : 31

מגישים :

איתי נוי 205702665
מורז פיינגולד 313608481
מתן ספירו 205731748
יפעת יוסקוביץ' 204736581

תוכן עניינים

5.....	טבלת המשתנים :	1.
5.....	תיאור המשתנים :	2.
6.....	תיאור קשרים בין משתנים :	3.
7.....	ניתוח תיאורי של המשתנים :	4.
14.....	פונקציית צפיפות והתפלגות מצטברת :	6.
15.....	ייצוג קשרים בעזרת תרשימים :	7.
18.....	טבלאות שכיחות :	8.
18.....	טבלאות חד ממדיות :	8.1
19.....	טבלאות דו ממדיות :	8.2
20.....	תקציר מנהלים :	9.
21.....	עיבוד מקדים :	10.
21.....	. הסרה של משתנים :	10.1
23.....	התאמת משתנים :	10.2
24.....	הגדרת משתני דמה :	10.3
24.....	הוספת משתני אינטראקציה :	10.4
27.....	התאמת המודל ובדיקת הנחות המודל :	11.
27.....	בחירת משתני המודל :	11.1
29.....	בדיקת הנחות המודל :	11.2
31.....	דוגמה לשימוש במודל הנבחר :	11.3
31.....	שיפור המודל :	12.
35.....	נספחים :	13.
35.....	נספח א' - תרשים קורלציה	13.1
35.....	נספח ב' - דוגמה לקטע הקוד של ניתוח תיאורי של המשתנים	13.2
36.....	נספח ג' - פלט הטבלאות החד מימדיות	13.3
36.....	נספח ד' - פלט הטבלאות הדו מימדיות	13.4
37.....	נספח ה' - Summaries של משתנים שמועמדים להסרה	13.5
39.....	נספח ו' - תוצאות האלגוריתמים	13.6
42.....	נספח ז' - בדיקת הנחות המודל :	13.7
42.....	נספח ח' - שימוש המודל	13.8
43.....	נספח ט' - שיפור המודל	13.9
45.....	נספח י' - קוד המודל	13.9.1

רשימת טבלאות

5	טבלה 1 - פירוט המשתנים
7	טבלה 2 - ניתוח המשתנים הרציפים
9	טבלה 3 - ניתוח המשתנה הקטגוריאל "משטר"
10	טבלה 4 - ניתוח המשתנה הקטגוריאל "דת"
11	טבלה 5 - ניתוח חריגים
18	טבלה 6 - שכיחות מס' ילדים ממוצע
18	טבלה 7 - שכיחות דרגת אושר ממוצעת
19	טבלה 8 - שכיחות גיל בעת נישואין ומס' ילדים ממוצע
19	טבלה 9 - שכיחות שעות עבודה ושכר
21	טבלה 10 - נתונים סטיסטיים על המשתנים רציפים

רשימת גרפים

11	איור 1 - Boxplot Age.at.1st.birth
11	איור 2 - Boxplot Days.at.home.for.labor
11	איור 3 - Boxplot Rate.of.happiness
12	איור 4 - Boxplot-Yrs.of.education
12	איור 5 - Boxplot divorce.rates
12	איור 6 - Boxplot Life.expectancy
13	איור 7 - Boxplot Hrs.of.work
13	איור 8 - Boxplot Avg.marriage.age
13	איור 9 - Boxplot Avg.num.of.kids
13	איור 10 - Boxplot Avg.num.of.kids
14	איור 11 - Density Days.at.home.for.labor
14	איור 12 - Accumulative Days.at.home.for.labor
14	איור 13 - Density Avg.num.of.kids
14	איור 14 - Accumulative Avg.num.of.kids
15	איור 15 - Density Life.expectancy
15	איור 16 - Accumulative Life.expectancy
15	איור 17 - Scatterplot happiness~life.expenctacy
16	איור 18 - Scatterplot education~avg.num.of.kids
16	איור 19 - Scatter plot marriage~divorce
17	איור 20 - Scatterplot wage~education
17	איור 21 - Scatterplot wage~religion
18	איור 22 - Histogram Avg.num.of.kids
18	איור 23 - Histogram Rate.of.happiness
21	איור 24 - Scatterplot Y~Hrs.of.work
21	איור 25 - Scatterplot Y~Divorce.rates
21	איור 26 - Scatterplot Y~Days.at.home.for.labor
22	איור 27 - Scatterplot Y~religion
22	איור 28 - Scatterplot Y~regime
23	איור 29 - Scatterplot Y~regime
23	איור 30 - Summary(Y~regime)
24	איור 31 - Summary(Y~Rate.of.happiness)
24	איור 32 - Scatterplot marraige~regime~Y
25	איור 33 - Scatterplot education~regime~Y
25	איור 34 - Scatterplot Life.expectancy~regime~Y
25	איור 35 - Summary for previous scatterplot

26.....	Scatterplot happiness~regime~Y	- 36	איור 36
26.....	Scatterplot wage~regime~Y	- 37	איור 37
26.....	Scatterplot labor~regime~Y	- 38	איור 38
27.....	Scatterplot kids~regime~Y	- 39	איור 39
29.....	Summary of stepwise	- 40	איור 40
29.....	Scatterplot Yhat~FixedErrors	- 41	איור 41
30.....	Ftest	- 42	איור 42
30.....	Histogram of residuals	- 43	איור 43
30.....	QQplot of errors	- 44	איור 44
30.....	KS test	- 45	איור 45
31.....	Data for UAE	- 46	איור 46
31.....	Boxcox	- 47	איור 47
32.....	QQplot before Bboxcox	- 48	איור 48
32.....	QQplot after Boxcox	- 49	איור 49
32.....	QQplot before Boxcox	- 50	איור 50
32.....	QQplot after Boxcox	- 51	איור 51
32.....	New KS test	- 52	איור 52
33.....	QQplot for 1st trans	- 53	איור 53
33.....	QQplot for 2 nd trans	- 54	איור 54
33.....	Histogram for 1st trans	- 55	איור 55
33.....	Histogram for 2nd trans	- 56	איור 56
34.....	Final scatterplot Yhat~fixed errors	- 57	איור 57

1. טבלת המשתנים:
טבלה 1- פירוט המשתנים

סוג המשתנה - מוסבר/מסביר	סימון	יחידת מידה	סוג המשתנה – רציף / קטגוריאלי	הסבר קצר על המשתנה
מוסבר	Y	שנים	רציף	גיל לידה ראשונה
מסביר	X ₁	ימים	רציף	ימי חופשת לידה בתשלום
מסביר	X ₂	חסר יחידות	רציף	דרגת אושר
מסביר	X ₃	שנים	רציף	שנות השכלה לאשה
מסביר	X ₄	אחוז	רציף	אחוזי גירושין
מסביר	X ₅	שנים	רציף	תוחלת חיים ממוצעת לנשים
מסביר	X ₆	שעות	רציף	ממוצע שעות עבודה
מסביר	X ₇	שנים	רציף	גיל נישואים ממוצע לנשים
מסביר	X ₈	חסר יחידות	רציף	מספר ילדים ממוצע
מסביר	X ₉	דולר	רציף	שכר ממוצע נשים
מסביר	X ₁₀	חסר יחידות	קטגוריאלי	דת שולטת במדינה
מסביר	X ₁₁	חסר יחידות	קטגוריאלי	משטר שולט במדינה

2. תיאור המשתנים:

המשתנה המוסבר:

Y - גיל ממוצע ללידה ראשונה. משתנה זה נבחר להיות המשתנה המוסבר כי לדעתנו גורמים רבים עשויים להשפיע על ערכו. בין הגורמים שחקרנו קיימים גורמים כלכליים, חברתיים, דתיים, מדיניים ועוד.

המשתנים המסבירים:

משתנים רציפים:

X₁ - ימי חופשת לידה בתשלום לאישה במדינה: משתנה זה מציג את כמות הימים בהם תוכל האישה להיעדר בשל לידה ותקבל עליהם שכר. אנו מניחים כי ככל שימי חופשת הלידה רבים יותר, כך הגיל הממוצע ללידה ראשונה יהיה נמוך יותר. ניתן להניח כי ימי חופשה מועטים יגרמו לאישה להרגיש צורך לחסוך כסף, טרם כניסתה להיריון, כך שתוכל להבטיח יציבות כלכלית עבור תקופה זו, ולכן הלידה תתעכב.

X₂ - דרגת אושר למדינה: משתנה זה מציג את הדירוג של רמת האושר במדינה. אנו מניחים כי ככל שדרגת האושר גבוהה יותר, כך הגיל הממוצע ללידה ראשונה יהיה נמוך יותר. אנו רואים בדרגת האושר סממן להשקפה אופטימית, הגורמת לתהליך לידת הילד להיתפס כדבר חיובי ובכך מפחיתה את יחס האישה לקשיים ולחששות שעלולים להיווצר, עקב כך האישה לא תחשוש ולא תדחה את לידתה הראשונה.

X₃ - שנות השכלה לאשה למדינה: משתנה זה מציג את משך הזמן הממוצע של נשים במערכת החינוך ובאקדמיה במדינה. אנו מניחים כי ככל ששנות ההשכלה רבות יותר, כך הגיל הממוצע ללידה ראשונה יהיה גבוה יותר, זאת משום שמערכות אלו דורשות תשומת לב, השקעה וזמן לא מוגבל מהאשה, מה שעלול למנוע ממנה להקים משפחה ולטפל בילד.

X₄ - אחוזי גירושין למדינה: משתנה זה מציג את ההסתברות לאישה לגירושין. אנו מניחים כי ככל שאחוז הגירושין גבוה יותר, כך הגיל הממוצע ללידה ראשונה יהיה גבוה יותר, כיוון והאישה תרגיש פחות ביטחון בקשר וכתוצאה מכך תעכב את הקמת התא המשפחתי.

X₅ - תוחלת החיים לנשים במדינה: משתנה זה מגדיר את הגיל הממוצע אליו מגיעות נשים. אנחנו מניחים כי ככל שתוחלת החיים לנשים נמוכה יותר, כך הגיל הממוצע ללידה ראשונה יהיה נמוך יותר, זאת משום שתוחלת חיים נמוכה יותר תחזק את רצונה של האישה בשלב מוקדם יותר ליצירת משפחה. ידוע כי זהו רצון הקיים בטבע לצורך המשכיות.

X₆ - ממוצע שעות עבודה שבועיות לאישה למדינה: משתנה זה מציג את מספר השעות הממוצע שאישה משקיעה בעבודתה בשבוע, לפי מדינה. אנו מניחים כי ככל שממוצע שעות העבודה השבועיות גבוה יותר, כך הגיל הממוצע ללידה ראשונה יהיה גבוה יותר. זאת משום ששעות עבודה רבות יותר עלולות להשפיע על מידת הזמן הפנוי של האישה ויכולותיה לעמוד בהקמת משפחה ולטפל בילד.

X₇ - גיל נישואין ממוצע לנשים במדינה: משתנה זה מציג את גיל הנישואין הממוצע במדינה. אנו מניחים כי ככל שגיל הנישואין הממוצע נמוך יותר, כך הגיל הממוצע ללידה ראשונה יהיה נמוך יותר. אנו רואים בהקמת משפחה ולידה כצעד המגיע בצמוד לחתונה.

X₈ - מספר ילדים ממוצע במדינה: משתנה זה מציג את מספר הילדים הממוצע לאישה במדינה. אנו מניחים כי ככל שמספר הילדים גבוה יותר, כך הגיל הממוצע ללידה ראשונה יהיה נמוך יותר, זאת כיוון שנתון זה מאיץ את האישה להתחיל לבנות משפחה, בהתאם לנורמות המצויות בסביבתה, מתוך הבנה שעליה להתחשב בשנות הפוריות.

X₉ - שכר ממוצע שנתי לנשים במדינה: השכר הממוצע של נשים משקף את הערך השנתי בדולרים. אנחנו מניחים כי ככל ששכרה הממוצע של אישה גבוה יותר, כך הגיל הממוצע ללידה ראשונה יהיה גבוה יותר. זאת כיוון ששכר ממוצע גבוה משפיע על מידת ההשקעה והזמן הדרוש מאישה במשק. כתוצאה מכך, האישה תהיה פחות פנויה להקמת תא משפחתי ותדחה את רצונותיה ללידה ראשונה.

משתנים קטגוריאליים:

X₁₀ - דת שולטת במדינה: משתנה זה מציג את הדת המרכזית הקיימת במדינה אשר ככל הנראה משפיעה על נורמות המדינה. ישנן דתות בעלות אופי מסורתי אשר מעודדות ילודה בגילאים צעירים ושמות במרכז החיים את חיי המשפחה וגידול הילדים, וישנן דתות שתפיסתן הדוקה פחות. זהו משתנה קטגוריאלי, המסומן באופן הבא: 1=בודהיזם, 2=נצרות, 3=הינדואיזם, 4=חסידי דת, 5=יהדות, 6=אסלאם.

X₁₁ - משטר שולט במדינה: למשטר במדינה קיימת השפעה גדולה על אופי המדינה וחוקיה ובכך משפיע על התנהלות אזרחיה. אנו מניחים כי במדינות בהן המשטר נוקשה, מעמד האישה יהיה נמוך יותר ועל כן מרכז חייה יהיה משפחתה, מה שיגרום לגיל ממוצע לידה ראשונה להיות נמוך יותר. בנוסף, חיי האזרחים יהיו פחות גמישים מה שתומך בהנחתנו, כיוון שמסלול חייהם ברור וקבוע עבורם.

זהו משתנה קטגוריאלי, המסומן באופן הבא: 1=דמוקרטיה, 2=מפלגה דומיננטית, 3=כבוש, 4=ממשל צבאי, 5=מונרכיה, 6=חד מפלגתי, 7=דיקטטורה, 8=ארעי.

3. תיאור קשרים בין משתנים:

מטריצת קורלציה בנספח א'

קשרים סיבתיים:

1. **דרגת האושר X₂ לשכר הממוצע לאישה X₉:** לדעתנו ככל שהשכר של הפרט גבוה יותר כך רמת האושר שלו תעלה, מכיוון שבשכר גבוה יותר איכות החיים משתפרת מה שמוביל לאושר גדול יותר. כמו כן ככל שהשכר גבוה יותר, רמת הלחץ שהפרט חווה ביום-יום קטנה יותר מכיוון שסביר שיהיו לו פחות דאגות כלכליות.

2. **שיעור הגירושין X_4 לשעות העבודה X_6** : אנו מעריכים שככל שמס' שעות העבודה עולה, כך גם עולה שיעור הגירושין. זאת מכיוון שלהערכתנו, זוג ממוצע עשוי לחוות קשיים במידה ואחד או שני בני הזוג עובדים שעות מרובות.
3. **שיעור הגירושין X_4 לשכר X_9** : אנו מעריכים שככל שהשכר עולה, כך גם יורד שיעור הגירושין. זאת מכיוון להערכתנו, זוג ממוצע שיחווה פחות קשיים כלכליים, יחווה פחות לחצים ועל כן יצליח לשמור על קשר בריא ויציב יותר.
4. **שלטון במדינה X_{11} וזת שולטת במדינה X_{10}** : להערכתנו, מדינות בעלות שלטון מונרכי בדור"כ מאופיינות בדתות שמרניות, מכיוון ששלטון מסוג זה נוטה לערב את הדת בהחלטות מדיניות ובאופי השלטון.

קשרים מדגמיים:

1. **מספר ממוצע של ילדים X_8 לתוחלת חיים X_5** - ניתן לראות כי מקדם המתאם בין שני המשתנים הינו $\rho = -0.819$ מה שמצביע על קשר מדגמי שלילי בין המשתנים, כלומר ככל שתוחלת החיים קטנה יותר, כך מספר ממוצע של ילדים יהיה גדול יותר. מכיוון שמקדם המתאם בערכו המוחלט גדול, הקשר המדגמי ייחשב כחזק. עם זאת, נוכל להסיק כי הקשר הסיבתי בין המשתנים דווקא הפוך, כי סביר שככל שתוחלת החיים גדולה יותר, המספר הממוצע של הילדים יהיה יותר גדול. לכן נסיק כי קיימים משתנים נוספים שהשפיעו על הקשר בצורה עקיפה, וכאשר נכניס אותם למשוואה הקשר הסיבתי בין המספר הממוצע של הילדים לתוחלת חיים יראה מובהקות גבוהה יותר.
2. **מספר שנות השכלה X_3 וגיל נישואים ממוצע לנשים X_7** - ניתן לראות כי מקדם המתאם בין שני המשתנים הינו $\rho = 0.7148$ מה שמצביע על קשר מדגמי חיובי בין המשתנים, כלומר ככל שמספר שנות הלימוד גדול יותר, כך גדל גם גיל הנישואין הממוצע. נטען כי הקשר מדגמי בלבד, שכן לא ניתן לקבוע גורם סיבתי בין שני המשתנים. הקשר המדגמי הנוצר עלול לנבוע מהמדגם הספציפי שלקחנו, ומאקראיות הנתונים, בנוסף ייתכן ונובע מקשר עקיף שנובע מפרמטרים אחרים, למשל פרמטר המצביע האם המדינה מפותחת או לא, שכן סביר להניח שבמדינה מפותחת מספר שנות הלימוד יהיה גבוה יותר.

4. ניתוח תיאורי של המשתנים: קטע הקוד הרלוונטי מצורף בנספח ב'

משתנים רציפים:

טבלה 2- ניתוח המשתנים הרציפים

הפלט	ניתוח נתונים	משתנה
Mean : 24.87 Median : 24.00 SD= 4.10 Interquartile range= [21.27,28.62] Skewness= 0.19	במדד זה ניתן לראות שהממוצע קרוב מאד לחציון ולכן ניתן להסיק שההתפלגות סימטרית. כמו כן הנתון של הצידוד מאשרר את זה שמדובר שהתפלגות סימטרית מכיוון שהזנב הימני קטן.	גיל בלידה (Y)ראשונה

Mean: 108.03 Median : 98.00 SD= 54.94 Interquartile range = [84,120] Skewness= 3.03	למשתנה מסביר זה יש סטיית תקן גדולה מאד, מה שמתאים לידע המקדים שלנו שלכל מדינה בעולם יש זכות לקבוע את ענייני הפנים שלה באופן עצמאי, ואין נוהל אחיד בעניין זה. ניתן לראות שיש זנב ימני גדול מאד, ולכן אנו מסיקים שההתפלגות אינה סימטרית. ממוצע הרבעונים גדול מהחציון, מה שאמור לגרום לזנב ימני ואכן זה המקרה.	ימי חופשת לידה (X_1) בתשלום
Mean : 5.47 Median : 5.44 SD= 1.12 Interquartile range = [4.54, 6.25] Skewness= -0.01	במשתנה מסביר זה הממוצע צמוד לחציון ובנוסף הצידוד כמעט אפסי- לכן ברור לנו שמדובר בהתפלגות סימטרית.	דרגת אושר (X_2)
Mean: 7.73 Median: 8.3 SD= 3.51 Interquartile range = [5.05, 10.52] Skewness= -0.32	ניתן לראות שישנה סטיית תקן גדולה אך עם זאת צידוד יחסית קטן- זנב שמאלי. כמו כן הממוצע יחסית קרוב לחציון ולכן הנתונים מציגים התפלגות כמעט סימטרית, עם סטייה יחסית קטנה כאמור.	שנות השכלה לאשה (X_3)
Mean: 41.86 Median: 44.23 SD= 8.95 Interquartile range = [39.74, 47.18] Skewness= -1.02	במשתנה זה אמנם הממוצע יחסית קרוב לחציון, אך ניתן לראות שקיים זנב שמאלי משמעותי ולכן נסיק שההתפלגות לא סימטרית. ממוצע הרבעונים קטן מהחציון, מה שאמור לגרום לזנב שמאלי ואכן זה המקרה.	אחוזי גירושין (X_4)
Mean: 70.48 Median: 71.65 SD= 7.76 Interquartile range = [65.21, 76.80] Skewness= -0.55	להתפלגות משתנה זה קיים זנב שמאלי יחסית קטן מה שמעיד על התפלגות לא סימטרית באופן מלא. כמו כן, הממוצע יחסית קרוב לחציון אבל כצפוי- לא קרוב מספיק כדי לראות התפלגות סימטרית	תוחלת חיים נשים (X_5)
Mean: 43.28 Median: 40 SD= 4.08 Interquartile range = [40, 48] Skewness= 0.23	להתפלגות משתנה זה קיים זנב ימני יחסית קטן מה שמעיד על התפלגות לא סימטרית באופן מלא. כמו כן, הממוצע יחסית קרוב לחציון אבל כצפוי- לא קרוב מספיק כדי לראות התפלגות סימטרית	שעות עבודה (X_6)
Mean: 24.24 Median: 24.1 SD= 4.93	ניתן לראות שהממוצע כמעט וזהה לחציון ושהצידוד כמעט אפסי (זנב ימני קטן)- ניתן להסיק בוודאות שההתפלגות סימטרית	גיל נישואים ממוצע לנשים (X_7)

Interquartile range = [20.15, 28.85] Skewness= 0.05		
Mean: 2.55 Median: 2.1 SD= 1.19 Interquartile range = [1.70, 3.25] Skewness= 1.15	במשתנה זה קיימת אי סימטריה בשל הזנב הימני הגדול. נסיק שההבדל בין הממוצע לחציון גדול, על אף שמדובר לכאורה במספרים קרובים. ממוצע הרבעונים גדול מהחציון, מה שאמור לגרום לזנב ימני ואכן זה המקרה.	ממוצע ילדים (X_8)
Mean: 14156.99 Median: 8901.84 SD= 14548.46 Interquartile range = [2915.33, 21090.55] Skewness= 1.42	למשתנה זה יש צידוד חיובי גדול- ז"א זנב ימני גדול. ניתן גם לראות שהממוצע שונה מאוד מהחציון, מה שמתאים לידע המקדים שלנו בנוגע להבדל בין שני המדדים: הנתונים מקבלים משמעות שונה לגמרי כאשר בוחנים אותם לפני כל מדד, ודוגמה טובה לכך היא מדדים מדיניים כאלה. גם בישראל הפער ביניהם גדול מאוד ולכן על מקבלי ההחלטות להתייחס לשניהם. ממוצע הרבעונים גדול מהחציון, מה שאמור לגרום לזנב ימני ואכן זה המקרה.	שכר (X_9)

משתנה קטגוריאל- משטר X_{11} :

טבלה 3- ניתוח המשתנה הקטגוריאל "משטר"

Skewness	Interquartile range	S.D	Median	Mean	n	משטר	אינדקס
-0.08	[21.3,29]	4.2	26	25.40	88	Democracy	1
0.43	[20,28.6]	5.1	22.9	24.48	9	Dominant Party	2
0	[27.775,28.725]	1.34	28.25	28.25	2	Foreign/Occupied	3
-0.62	[22.275,23.075]	1.02	22.85	22.5	4	Military	4
1.27	[23,24.2]	2.09	23.5	24.075	8	Monarchy	5
-0.1	[21.2,23]	1.39	22.3	22.12	5	Party-Personal	6
0.51	[20.35,23.5]	3.18	21.3	22.25	11	Personal Dictatorship	7
NA	[32.6,32.6]	NA	32.6	32.6	1	Provisional - Civilian	8

משטר דמוקרטי: החציון והממוצע כמעט זהים וניתן לראות כי קיים זנב שמאלי מזערי ע"פ ה-Skewness השלילי ובעל ערך מוחלט נמוך מאוד, דבר זה מדגיש על הסימטריה של פיזור הנתונים. טווח הגילאים נראה רחב, מה שסביר למשטר ליברלי בעל דעות מגוונות בשונה ממשטרים מקובעים אחרים. משטר דיקטטורי: ניתן לראות זנב ימני שכן החציון נמוך מהממוצע, ו Skewness חיובי משטר

משטר מפלגה דומיננטית: ניתן לראות זנב ימני קטן על פי ה Skewness החיובי, וכן ממוצע גבוה מהחציון. נשים לב גם לטווח הגדול בגילאים.

יתר הקטגוריות: שאר הקטגוריות בעלות מספר תצפיות קטן ועל כן לא נסיק על התנהגות סטטיסטית עבור כל אחת מהן.

משתנה קטגוריאל - דת X₁₀:

טבלה 4- ניתוח המשתנה הקטגוריאל "דת"

אינדקס	Religion	n	Mean	Median	S.D	Interquartile range	Skewness
1	Buddhist	7	23.82	23	3.38	[21.95, 24.45]	0.89
2	Christian	81	25.25	26.3	4.09	[21.2, 28.9]	-0.08
3	Hindu	3	25.02	23	5.52	[21.9, 27.13]	0.31
4	Irreligion	3	23.82	23	3.38	[21.95, 24.45]	0.89
5	Jewish	1	27.6	27.6	27.6	[27.6, 27.6]	.NA
6	Muslim	33	23.68	23	4.10	[21.9, 24.6]	0.92

דת הנצרות: קיים הבדל בין הממוצע והחציון אך נראה שהצידוד כמעט לא קיים, מדובר בזנב שמאלי קטן מאוד. ניתן לראות כי טווח הגילאים מתפרש על טווח ערכים גדול. ניתן להסביר זאת על ידי ההנחה שמדינה שהדת הנוצרית בה הינה השולטת, יכולה להיות מפותחת או מתפתחת, מה שמשפיע בהתאמה על גילאי ילודה ראשונה.

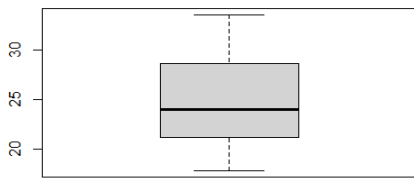
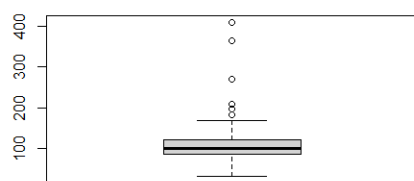
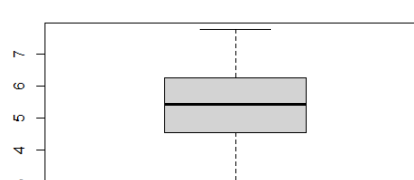
דת האסלאם: הממוצע והחציון אמנם קרובים אך קיים זנב ימני משמעותי שמצביע על חוסר סימטריה בהתפלגות ילודת ילד ראשון. הנתונים מתאימים למידע שיש לנו מהמציאות- ברוב המדינות המוסלמיות נשים יולדות בגיל יחסית צעיר, אך נסיק כי קיים פיזור רחב של מדינות בהן הגיל במעט גדול יותר.

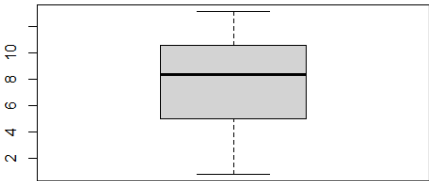
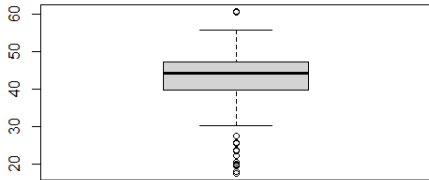
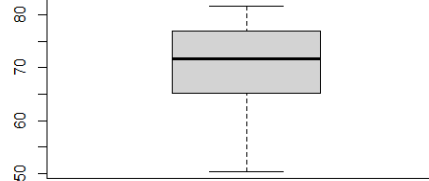
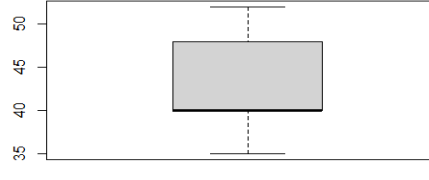
יתר הדתות: שאר הקטגוריות בעלות מספר תצפיות קטן ועל כן לא נסיק על התנהגות סטטיסטית עבור כל אחת מהן.

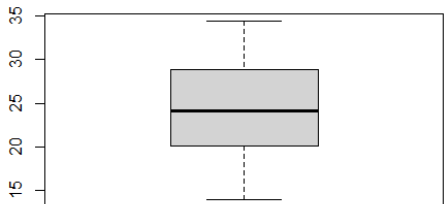
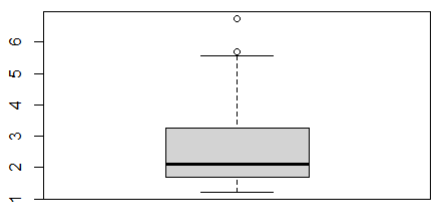
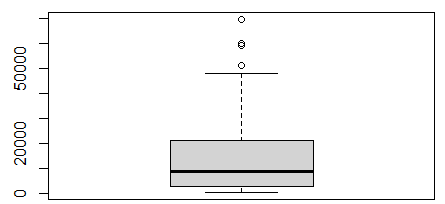
5. ניתוח חריגים:

טבלה 5- ניתוח חריגים

אנו מציגים את היישום בקוד [בנספח ח'](#).

Boxplot	ניתוח	המשתנה
<p>Age.at.1st.birth</p>  <p>Boxplot Age.at.1st.birth - איור 1</p>	<p>אין נתונים חריגים. המדידות מתרכזות מסביב לחציון.</p>	<p>גיל ממוצע ללידה ראשונה (Y)</p>
<p>Days.at.home.for.labor</p>  <p>Boxplot Days.at.home.for.labor - איור 2</p>	<p>קיימת מדינה אחת בעלת כמות נמוכה באופן חריג של ימי חופשת לידה בתשלום- אפגניסטן המאופיינת במעמד אישה נמוך ועוני רב דבר המשפיע על חקיקה זו. מנגד המדינות בעלות כמות הימים החריגה באופן גבוה בעלות מאפיינים הפוכים לאפגניסטן.</p> <p>ביניהן: בולגריה, סן-מרינו, ואלבניה.</p> <p>נבחר להשאיר נתונים חריגים אלה כיוון שאנו מניחים כי ימי חופשת לידה רבים יעודדו לידה מוקדמת.</p>	<p>ימי חופשת לידה בתשלום (X_1)</p>
<p>Rate.of.happiness</p>  <p>Boxplot Rate.of.happiness - איור 3</p>	<p>אין נתונים חריגים. המדידות מתרכזות מסביב לחציון.</p>	<p>דרגת אושר (X_2)</p>

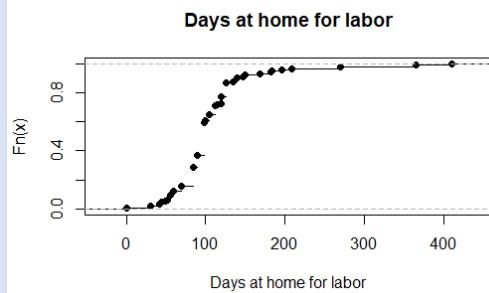
<p style="text-align: center;">Yrs.of.education</p>  <p style="text-align: center;">Boxplot-Yrs.of.education - איור 4</p>	<p>שנות השכלה לאשה (X_3)</p> <p>אין נתונים חריגים.</p>	
<p style="text-align: center;">Divorce.rates</p>  <p style="text-align: center;">Boxplot divorce.rates - איור 5</p>	<p>אחוזי גירושין (X_4)</p> <p>ניתן לראות כי קיימות מספר מדינות בהן אחוז הגירושים הינו נמוך באופן חריג, ומדינה אחת בעלת אחוז גירושין גבוה במיוחד - טורקיה. ניתן להסביר אחוז זה ע"י השינוי שהתרחש במדינה באורח חיי האזרחים, והקלות שבוצעו להליך הגירושין. מנגד בין המדינות בעלות האחוז הנמוך נמצאות - איראן, צרפת, ג'מייקה. באיראן למשל בשביל שאישה תפתח בגירושין עליה להוכיח כי המשך הנישואין יגרום לה לנזק.</p> <p>נבחר להשאיר נתונים אלה כיוון והם מייצגים את מידת נוקשות המדיניות בה נוקטת כל ממשלה.</p> <p>נרצה לבחון את הקשר בין משתנה זה למשתנה המוסבר.</p>	
<p style="text-align: center;">Life.expectancy</p>  <p style="text-align: center;">Boxplot Life.expectancy - איור 6</p>	<p>תוחלת חיים ממוצעת לנשים (X_5)</p> <p>אין נתונים חריגים, יש אסימטריה ימינה כיוון ואנו מוגבלים מצד שמאל.</p>	
<p style="text-align: center;">Hrs.of.work</p> 	<p>ממוצע שעות עבודה (X_6)</p> <p>אין נתונים חריגים. ניתן לראות כי החציון בעל הערך של האחוזון ה-25, הדבר נובע מכך שבמרבית המדינות קיימת הגבלה דומה על מספר שעות העבודה לעובד בשבוע.</p>	

<p>Boxplot Hrs.of.work - 7 איור</p>		
<p>Avg.marriage.age</p>  <p>Boxplot Avg.marriage.age - 8 איור</p>	<p>אין נתונים חריגים. המדידות מתרכזות מסביב לחציון.</p>	<p>גיל נישואים ממוצע לנשים (X_7)</p>
<p>Avg.num.of.kids</p>  <p>Boxplot Avg.num.of.kids - 9 איור</p>	<p>קיימות שתי מדינות בהן מספר הילדים הממוצע הינו גבוה באופן חריג. ביניהן: ניגריה – שהינה המדינה הכי מאוכלסת באפריקה ולכן מתכתב עם חריגה זו. וסומליה אשר ייתכן כי בעלת יצר הישרדותי עקב תנאי המדינה הקשים.</p>	<p>מספר ילדים ממוצע (X_8)</p>
<p>Wage</p>  <p>Boxplot Avg.num.of.kids - 10 איור</p>	<p>קיימות מספר מדינות בהן השכר הממוצע גבוה שוויץ, נורבגיה, וסינגפור. באופן חריג. ביניהן: נבחר להשאיר חריגות אלה כיוון ומייצגות את עושר המדינה בצירוף למעמד הנשים בה, ובהתאם את השתכרות הנשים.</p> <p>ניתן לראות זנב שמאלי, היות וישנן יבשות רחבות המכילות מדינות עוני רבות.</p>	<p>שכר ממוצע נשים (X_9)</p>

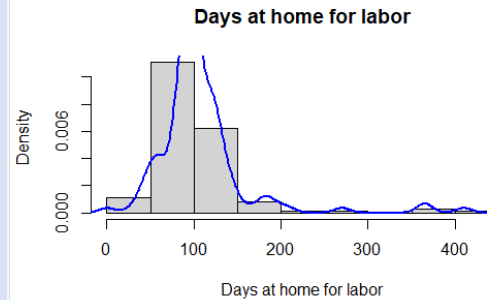
6. פונקציית צפיפות והתפלגות מצטברת:

אנו מציגים את היישום בקוד [בנספח ח'](#).

ימי חופשת לידה בתשלום (X_1)



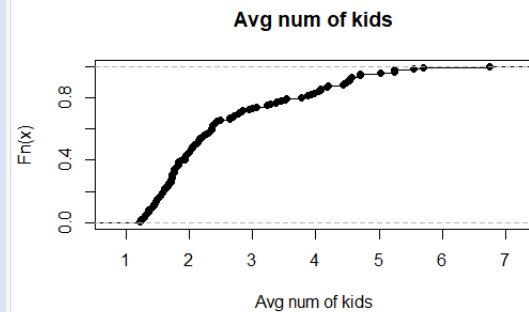
איור 12 - Accumulative Days.at.home.for.labor



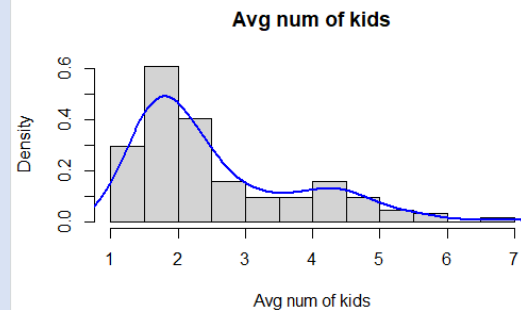
איור 11 - Density Days.at.home.for.labor

ניתן לראות עבור המשתנה המייצג את ימי החופשה, כי קיים זנב ימני המעיד על צפיפות אסימטרית ימנית. בנוסף, ניתן לראות כי עד שכר של כ-150\$ הגרף תלול מאוד, ולאחר מכן מתמתן ומתיישר כמעט לחלוטין, כך שההסתברות לקבל ימים רבים של חופשת בתשלום הולך וקטן. מרבית המדינות בעלות שכר ממוצע דומה המתפרש בין 50\$ ל-150\$ בשבוע, כאשר שיא פונקציית הצפיפות הינו בשכר של 100\$. ניתן לקשר טווח זה לגיל בו התינוק עדיין זקוק ליניקה מאמו לטובת התפתחותו.

מספר ילדים ממוצע (X_8)



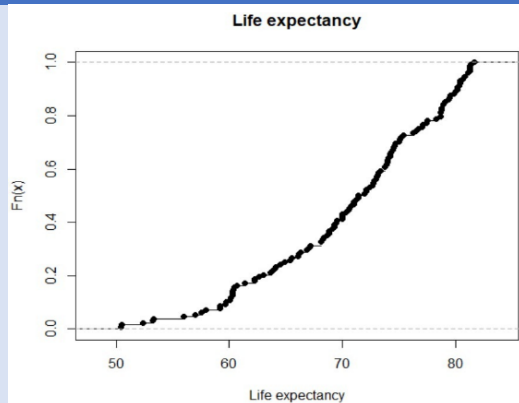
איור 14 - Accumulative Avg.num.of.kids



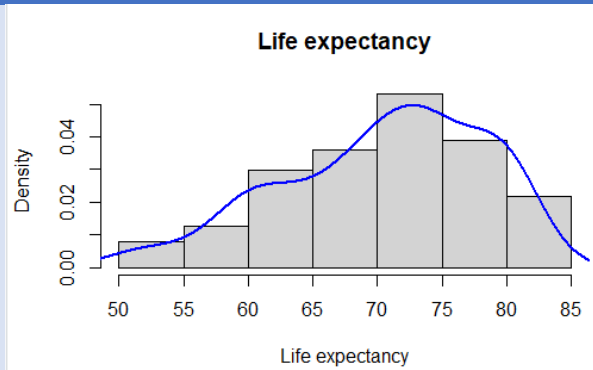
איור 13 - Density Avg.num.of.kids

ניתן לראות כי עבור המשתנה שמייצג את מספר הילדים הממוצע במדינה קיים זנב ימני, כלומר הצפיפות אסימטרית ימנית. שיא הצפיפות מתרחש סביב שני ילדים, כאשר מרבית הערכים נעים בין ילד אחד לשלושה ילדים ולכן סביב ערכים אלה גרף ההתפלגות המצטברת תלול. ניתן להסביר טווח זה ע"י הסתכלות על תמיכת הממשלות המוגבלת במספר הילדים למשפחה עקב הצפיפות והיכולת הכלכלית במדינה, וכתוצאה מכך אין יכולת כלכלית לנשים רבות להביא ילדים מעבר למספר המוגבל לתקציב.

תוחלת חיים ממוצעת לנשים (X_5)



איור 16 - Accumulative Life expectancy



איור 15 - Density Life expectancy

ניתן לראות עבור המשתנה המייצג את תוחלת החיים הממוצעת של אישה, כי גרף הצפיפות הינו יחסית פעמוני עם נטייה לזנב שמאלי. ניתן לייחס זנב זה לכך שערכי המשתנה מוגבלים משמאל. מרבית הערכים נמצאים סביב הגילאים 65 ל-80, כאשר שיא פונקציית הצפיפות מתרחש בין הגילאים 70 ל-75. ניתן לראות כי שיפוע פונקציית ההתפלגות המצטברת תואם למבנה הפעמוני, התפרשות זו ניתנת להסברה ע"י מגוון תנאי מחייה, רמת הרפואה והתברואה שהממשלות השונות מספקות.

7. ייצוג קשרים בעזרת תרשימים:

תרשים פיזור המציג את תוחלת חיים כתלות בדרגת האושר



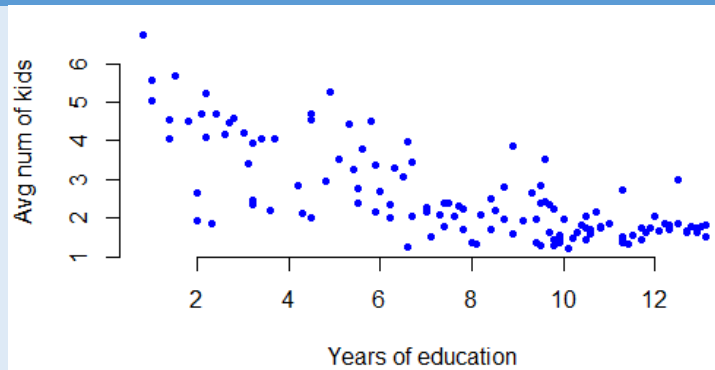
איור 17 - Scatterplot happiness~life.expency-17

ציר X: דרגת האושר (X_2)

ציר Y: תוחלת חיים (X_5)

בתרשים זה ניתן לראות מגמה המשקפת קשר חיובי בין המשתנים. ככל שדרגת האושר במדינה גבוה יותר כך תוחלת החיים של האישה מתארכת. קשר זה ניתן להסברה ע"י השפעת האושר של האדם על המערכת החיסונית והרצון שלו לחיות שנים רבות, המשפיעים באופן ישיר על תוחלת החיים. בנוסף, בהסתכלות על רמות האושר הגבוהות ניתן לראות באופן מובהק את הקשר החיובי לתוחלת החיים, המקבלת את ערכיה הגבוהים באופן הצפוף ביותר.

תרשים פיזור המציג את מספר הילדים הממוצע כתלות במספר שנות השכלה לאישה



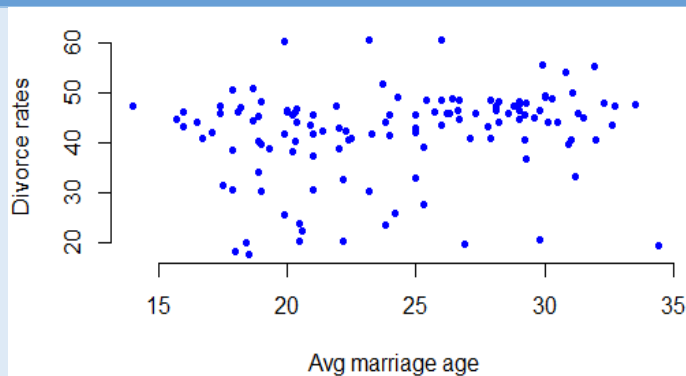
ציר X: שנות השכלה לאישה (X_3)

ציר Y: מספר ילדים ממוצע (X_8)

איור 18 - Scatterplot education~avg.num.of.kids

בתרשים זה ניתן לראות קשר שלילי בין המשתנים, כלומר ככל שאשה רוכשת שנות השכלה רבות יותר כך מספר הילדים שלה יפחת. ניתן להסביר זאת ע"י ההבנה כי השכלה דורשת השקעה רבה, דבר הגורר ילודה מועטה על מנת שתהיה יכולת השקעה סבירה לגידול כל ילד. עבור שנות השכלה מועטות ערכי מספרי הילדים מתפרשים על טווח רחב יותר, מאשר עבור שנות השכלה רבות.

תרשים פיזור המציג את אחוז הגירושין כתלות גיל ממוצע לנישואין



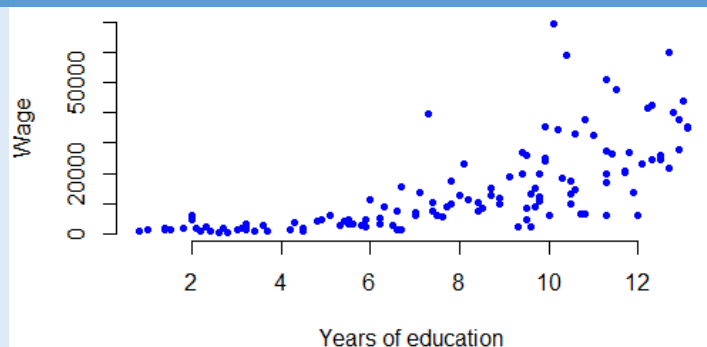
ציר X: גיל ממוצע לנישואין (X_7)

ציר Y: אחוז הגירושין (X_4)

איור 19 - Scatter plot marriage~divorce

בתרשים זה לא ניתן לזהות בבירור קשר בין המשתנים, על אף הנחתנו כי ככל שגיל הנישואין מוקדם יותר כך הסיכוי להתגרש גבוה יותר. אחוזי הגירושין מתרכזים סביב טווח הערכים 40%-50% ללא קשר מובהק לגיל ממוצע לנישואין. ניתן לראות פיזור בטווח רחב יותר סביב הגילאים הצעירים.

תרשים פיזור המציג את השכר הממוצע כתלות במספר שנות השכלה לאישה



איור 20 - Scatterplot wage~education

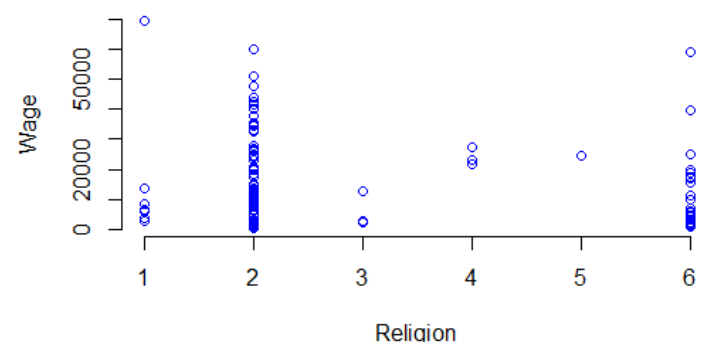
ציר X: שנות השכלה לאישה (X_3)

ציר Y: שכר ממוצע לאישה (X_9)

בתרשים זה קיים קשר חיובי בין המשתנים, כפי שהיינו מצפים, שנות ההשכלה לאישה משפיעות באופן חיובי על השתכרותה. ניתן להסביר זאת כיוון ושנות השכלה רבות מכוונות את האישה לתפקידים יוקרתיים יותר וגם מקנות ידע מקצועי המתבטא בתקרת שכר גבוהה יותר.

כמו כן, במידה ונוריד את התצפיות החריגות נראה שהקשר הינו אקספוננציאלי.

תרשים פיזור המציג את השכר הממוצע לאשה כתלות בדת



איור 21 - Scatterplot wage~religion

ציר X: דת (X_{10})

1=בדחזים

2=נצרות

3=הינדואיזם

4=חסרי דת

5=יהדות

6=אסלאם

ציר Y: שכר ממוצע לאשה (X_9)

בתרשים זה ניתן לראות כי מרבית הערכים מתרכזים בתחתית הגרף, כלומר בטווח משכורות נמוך-בינוני, ניתן להסביר זאת כיוון והשתכרות נשים נמוכה מהשתכרות גברים ברב המדינות, וקיימות במאגר הנתונים מדינות עוני רבות ולכן מרבית משכורות הנשים יהיו נמוכות. בנוסף, ניתן לראות כי ייתכן וקיים הבדל בין טווחי השכר שאשה מרוויחה כתלות בדת אליה משויכת, אך לא ניתן לומר באופן מובהק כיוון ואין מספיק נתונים לכל הדתות. ניתן להשוות את המדינות הנוצריות למוסלמיות מבחינת טווח השתכרות הנשים, כאשר אישה יכולה להגיע לשכר גבוה במדינות נוצריות רבות ומנגד קיימות מעט מדינות מוסלמיות בהן תוכל להגיע לרמות שכר גבוהות. ניתן להסביר זאת מכיוון שבמדינות מוסלמיות מעמד האישה נמוך יותר ועל כן נצפה ששכרה יהיה בהתאם. מבחינת הדתות הפחות נפוצות, ניתן לראות כי המדינות בהן הדת השולטת הינה בודהיזם והינדואיזם. שכר הנשים מתרכז בטווח נמוך יותר ממדינות חסרות דת שולטת, ויהודיות. ניתן להסביר זאת ע"י הסתכלות על אורח החיים המאופיין לדת אך לא ניתן להגיד באופן מובהק מכיוון שקיימות מעט מדינות המאופיינות בדתות אלה.

8. טבלאות שכיחות:

8.1. טבלאות חד ממדיות:

הפלט מוצג [בנספח ג'](#)

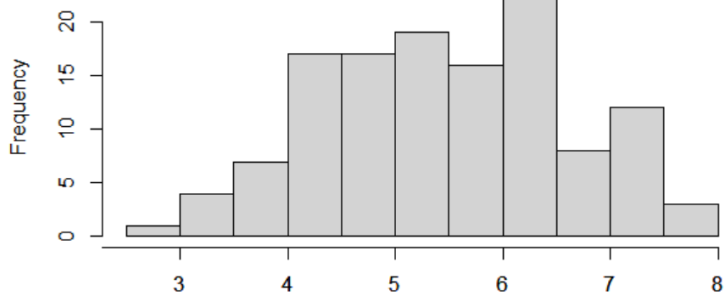
טבלה 6- שכיחות מס' ילדים ממוצע

טבלה 7 - שכיחות דרגת אושר ממוצעת

Rate of happiness	n	Percent
2 - 3	1	0.78%
3 - 4	11	8.59%
4 - 5	34	26.56%
5 - 6	35	27.34%
6 - 7	32	25.00%
7 - 8	15	11.72%

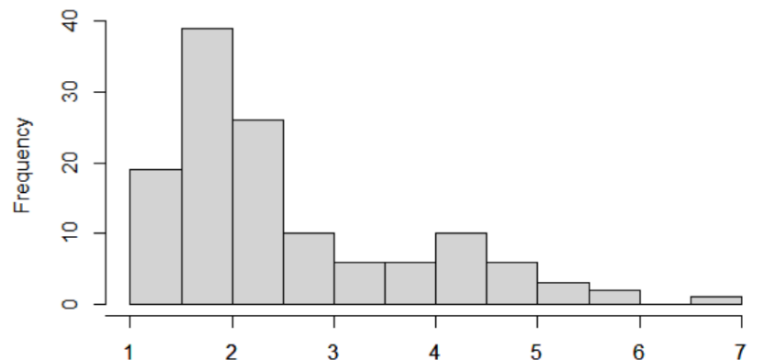
Avg num of kids	n	Percent
0-1	0	0.00%
1-2	58	45.31%
2-3	36	28.13%
3-4	12	9.38%
4-5	16	12.50%
5-6	5	3.91%
6-7	1	0.78%

Histogram of dataset\$Rate.of.happiness



איור 23- Histogram Rate.of.happiness

Histogram of dataset\$Avg.num.of.kids



איור 22- Histogram Avg.num.of.kids

מס' הילדים הממוצע X_8

ניתן לראות שהתפלגות מס' הילדים הממוצע אינה סימטרית, אלא בעלת זנב ימני גדול. ניתן לראות מהטבלה שברוב העולם מס' הילדים הממוצע הוא בטווח של 1-2. לאחר מכן קיים ריכוז משמעותי בטווח של 2-3 ובטווחים הבאים מדובר באחוזים קטנים יחסית.

דרגת האושר הממוצעת X_2

ההתפלגות של דרגת האושר דומה להתפלגות הנורמאלית כאשר מדובר בפעמון רחב. לפי הטבלה ניתן לראות פיזור בנתוני דרגת האושר, מה שמצביע על סטיית תקן גדולה. הגרף מזכיר התפלגות נורמאלית עם פעמון רחב, ז"א פיזור גדול של נתונים.

8.2. טבלאות דו ממדיות:

הפלט מוצג בנספח ד'

טבלה 8 - שכיחות גיל בעת נישואין ומס' ילדים ממוצע

Ranges	0-1	1-2	2-3	3-4	4-5	5-6	6-7
14-17	0	2	0	0	1	2	1
17-20	0	2	6	2	12	2	0
20-23	0	3	11	8	3	1	0
23-26	0	6	12	1	0	0	0
26-29	0	15	5	1	0	0	0
29-32	0	23	2	1	0	0	0
32-35	0	6	0	0	0	0	0

השורות - טווחי גיל האישה בעת הנישואין (X_7)

העמודות - טווח מס' הילדים הממוצע

ניתן לראות כי הריכוז הגדול של הנתונים נמצא בטווחים של 1-3 ילדים וגיל 20-32. הדבר לא מפתיע ומתאים לידע שלנו לגביי העולם המערבי. נשים לב שישנם גם מדינות בהן יש תוצאות קיצוניות יותר, כמו ילודת 4-5 ילדים בגילאי 17-20 (12 מדינות!).

טבלה 9 - שכיחות שעות עבודה ושכר

Ranges	0-10000	10000-20000	20000-30000	30000-40000	40000-50000	50000-60000	60000-70000
20-25	0	0	0	0	0	0	0
25-30	0	0	0	0	0	0	0
30-35	0	0	0	0	0	0	0
35-40	2	0	1	3	2	0	1
40-45	31	13	13	4	2	1	1
45-50	30	13	3	2	0	1	0
50-55	4	0	0	0	1	0	0

השורות - טווח שעות העבודה (X_6)

העמודות - טווח השכר (X_9)

ניתן לראות שהשכיחות הגבוהה ביותר מתקיימת במדינות בהן נשים עובדות 40-45 שעות חודשיות ושכרן השנתי הוא עד 10,000. כמו כן, ישנן נשים שבמדינות מסוימות משתכרות שכר גבוה יותר, כלומר מעל 40,000 אך מס' המדינות הללו קטן. בנוסף ניתן לראות שככל שטווח השכר עולה, שכיחות הנתונים קטנה. נשים לב שטווחי השכר הגבוהים יותר, לאו דווקא מתקיימים עבור מדינות בהן שעות העבודה גבוהות יותר.

פרויקט ברגרסיה לינארית - חלק ב'

9. תקציר מנהלים:

בפרויקט זה התנסו בבניית מודל רגרסיה לינארית מרובה על בסיס נתונים שבחנו ובעזרת תכנת ה-RStudio. בפרויקט ניתחנו את המשתנים, הקשרים ביניהם, הסרנו משתנים שאינם תורמים למודל והסקנו מסקנות בנוגע למשתני המודל והשפעתם על המשתנה המוסבר. בחרנו לבחון את הקשר בין 11 משתנים שונים לבין גיל לידה ראשונה ממוצע לנשים. המשתנים המסבירים שבחנו הם (עבור כל מדינה): ימי חופשת לידה בתשלום, דרגת אושר, שנות השכלה לאשה, אחוזי גירושין, תוחלת חיים ממוצעת לנשים, ממוצע שעות עבודה, גיל נישואין ממוצע לשים, מספר ילדים ממוצע, שכר ממוצע נשים, דת שולטת במדינה, משטר שולט במדינה.

תחילה, בחנו את הקשר בין המשתנים המסבירים לבין המשתנה המוסבר באמצעות טבלת פירסון אשר מציגה קורלציה בין המשתנים השונים. משתנים שלא עמדו בסף המתאם שהגדרנו, הועמדו להסרה. משתנים אשר רמת טיב ההתאמה שלהם הייתה נמוכה ומשתנים שהציגו גרף פיזור שאינו מעיד על קשר גבוה, הוסרו.

לאחר מכן, בדקנו את האינטראקציה של כל משתנה מסביר עם המשתנה הקטגוריאל "משטר שולט" במדינה, המשתנה הקטגוריאל שנוטר במודל, כדי לבחון את השפעתם על המשתנה המוסבר.

בשלב הבא, כדי למצוא את המודל הסופי עם המשתנים שמסבירים בצורה הטובה ביותר את המשתנה המוסבר ביצענו את האלגוריתמים "רגרסיה בצעדים", "רגרסיה לאחור" ו"רגרסיה לפנים" כאשר קריטריון הבחירה בין החלופות השונות הוא מזעור מדד AIC. בהתאם למבחנים הסטטיסטיים השונים, בחרנו להגדיר את המודל עם 3 משתנים רציפים והמשתנה הקטגוריאל "משטר".

לאחר קבלת המודל המתאים בדקנו את קיום הנחות המודל, ע"י בחינת גרפים וביצוע מבחנים סטטיסטיים מתאימים (מבחן F לשוויון שונויות, מבחן SW, תרשים היסטוגרמה, תרשים שאריות וכו').

גילינו שהנחות הלינאריות ושוויון השונויות מתקיימות, אך הנחת הנורמליות של השגיאות אינה מתקיימת, ולכן ביצענו Box-Cox על המודל באמצעותו בחנו את הטרנספורמציה המתאימה ביותר למשתנה המוסבר, ע"מ לאפשר למודל לעמוד בכלל הנחות. מבחני ההשערות והגרפים הראו כי אנו עדיין לא עומדים בהנחת הנורמאליות של השגיאות המתוקננות. לאחר מכן ביצענו טרנספורמציות מוכרות, וראינו כי טרנספורמציה $f(y) = y^{-0.5}$ עומדת בכל הנחות המודל. בנוסף, בחנו את האפשרות לבצע טרנספורמציה על המשתנים המסבירים, אך נוכחנו לגלות על פי גרפי הפיזור עם המשתנה המוסבר, כי אין אנו מזהים התנהגות של פונקציה מוכרת, ועל כן בחרנו שלא לבצע טרנספורמציה זו.

נותנו עם המודל הסופי הבא:

$$\widehat{y^{-0.5}} = 3.46758 + (-4.977042) * rgmUni + 7.642933 * rgmDict \\ + 0.008311 * x_1 + 0.108996 * x_5 + 0.533411 * x_7 + 0.23976 * x_7 * rgmUni \\ + (-380903) * x_7 * rgmDict$$

10. עיבוד מקדים:

10.1 . הסרה של משתנים:

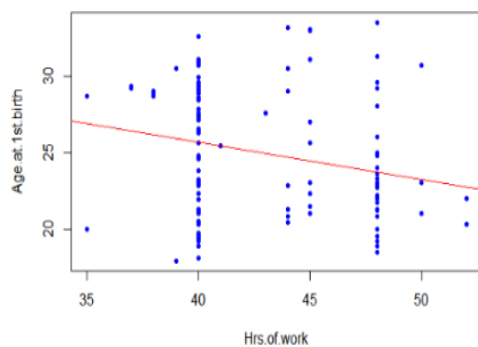
Wage	Avg.num.of.kids	Avg.marriage.age	Hrs.of.work	Life.expectancy..	Divorce.rates	Yrs.of.education.	Rate.of.happines	Days.at.home.for	Age.at.1st.birht	1	Age.at.1st.birht
0.610285852	-0.658940628	0.796631327	-0.241389121	0.687128163	0.163821177	0.664315261	0.525485017	0.296164697	0.296164697	1	Days.at.home.for
0.041831375	-0.255480084	0.22538795	-0.26857691	0.157252261	-0.099261694	0.216549685	0.038149928	0.038149928	0.525485017	1	Rate.of.happines
0.748126896	-0.632269603	0.626595709	-0.222073898	0.795222877	0.195944131	0.678195055	0.038149928	0.038149928	0.664315261	1	Yrs.of.education.
0.70070724	-0.742948091	0.71482867	-0.321717893	0.721734508	0.141412703	0.678195055	0.216549685	0.216549685	0.664315261	1	Divorce.rates
0.220128937	-0.126388659	0.203087095	-0.045755107	0.130505455	0.141412703	0.195944131	-0.099261694	0.195944131	0.687128163	1	Life.expectancy..
0.723732299	-0.819085135	0.718799705	-0.132005183	0.130505455	0.721734508	0.795222877	0.157252261	0.157252261	0.687128163	1	Hrs.of.work
-0.230532835	0.06649059	-0.330845089	0.130505455	-0.132005183	-0.045755107	-0.321717893	-0.222073898	-0.26857691	-0.241389121	1	Avg.marriage.age
0.660498106	-0.688698334	0.718799705	-0.330845089	0.718799705	0.203087095	0.71482867	0.626595709	0.22538795	0.796631327	1	Avg.num.of.kids
-0.577884164	0.06649059	-0.688698334	0.06649059	-0.819085135	-0.126388659	-0.742948091	-0.632269603	-0.255480084	-0.658940628	1	Wage
1	-0.577884164	0.660498106	-0.230532835	0.723732299	0.220128937	0.70070724	0.748126896	0.041831375	0.610285852	1	

משתנים רציפים:

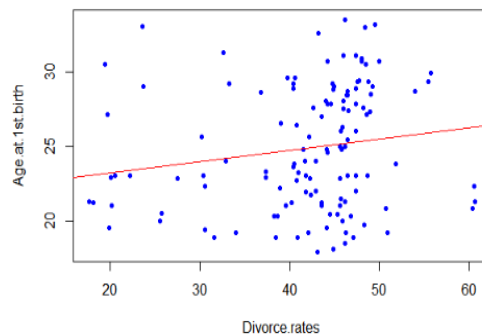
טבלה 10 - נתונים סטטיסטיים על המשתנים רציפים

Notation	Variable	Pearson	P-value
X1	Days at home for labor	0.2961647	0.000688
X2	Rate of happiness	0.525485	1.907e-10
X3	Yrs of education- woman	0.6643153	2.2e-16
X4	Divorce rates	0.1638212	0.06464
X5	Life expectancy- women	0.6871282	2.2e-16
X6	Hrs of work	-0.2413891	0.006052
X7	Avg marriage age- woman	0.7966313	2.2e-16
X8	Avg num of kids	-0.6589406	2.2e-16
X9	Wage	0.6102859	2.05e-14

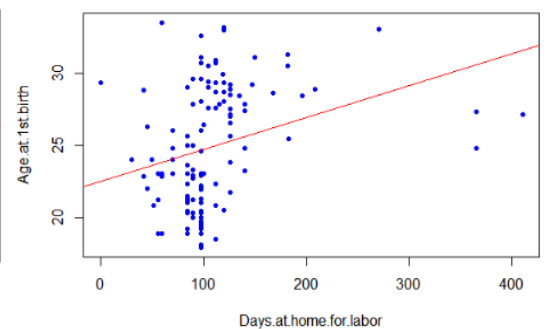
ניתן לראות כי קיימים מספר משתנים בעלי מתאם נמוך ואותם נסמן כמועמדים להסרה מהמודל. לאחר בחינת המתאם, נמשיך ונבחן את המשתנים החשודים להסרה ע"י הסתכלות על תרשימי הפיזור שלהם.



איור 24 - Scatterplot Y~Hrs.of.work



איור 25 - Scatterplot Y~Divorce.rates



איור 26 - Scatterplot Y~Days.at.home.for.labor

בתרשימי הפיזור נראה כי לא קיימים קשרים ליניאריים מובהקים. נמשיך לבחון משתנים אלה ע"י בחינת ה- R^2_{adj} וה-Pvalue, בעזרת מודלי הרגרסיה עבור כל משתנה מסביר אל מול המוסבר שנבחנו.

ה-Summaries מפורטים [בנספח ה'.](#)

X_6 : Hrs of work

בבדיקת הנתונים הסטטיסטיים ניתן לראות כי $R^2_{adj} = 0.05$, מה שמצביע על קשר לינארי נמוך מאוד בין משתנה זה למשתנה המוסבר (כפי שראינו בגרף הפיזור), ולמרות שה-Pvalue שלו נמוך (0.6%) וייתכן כי יש קשר בין המשתנים, איננו מסביר מספיק את המשתנה המוסבר ולכן נבחר להסיר את המשתנה.

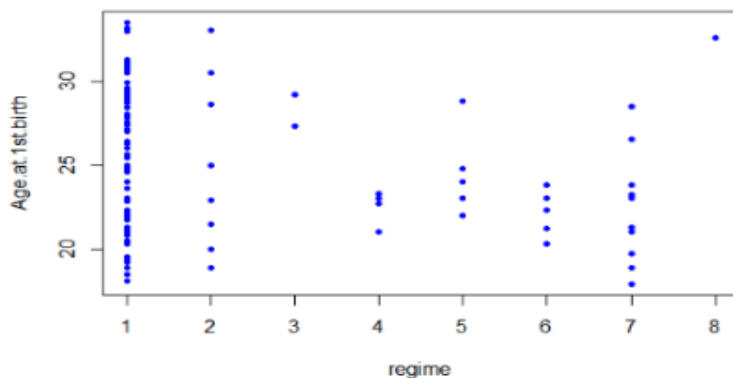
X_1 : Days at home for labor

לפי גרף הפיזור קשה לראות קשר לינארי. אך בבדיקת הנתונים הסטטיסטיים ניתן לראות כי ה- $R^2_{adj} = 0.08$, יחסית נמוך אך מייצג נתון גבוה מבין המשתנים שהועמדו להסרה. ה-Pvalue שלו נמוך מאוד (0.06%) וייתכן כי יש קשר בין המשתנים, ולכן נבחר שלא להסיר את המשתנה.

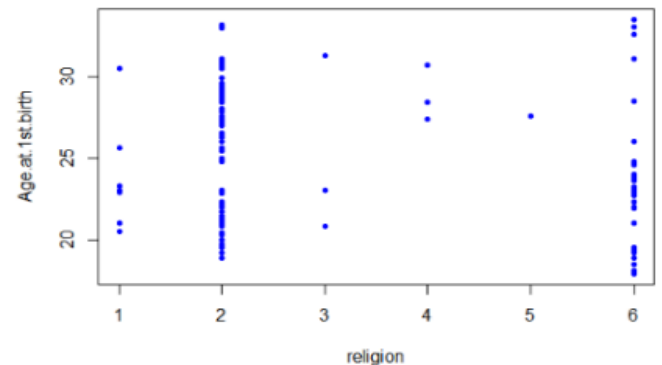
X_4 : Divorce rates

גרף המשתנה בעל פיזור רב ולכן סביר שלא קיים קשר לינארי. בבדיקת הנתונים הסטטיסטיים התגלה כי ה- $R^2_{adj} = 0.019$. הערכים הללו מזעריים (אנו שואפים ל-R גבוה שקרוב ל-1) ובנוסף ה-Pvalue שלו גבוה (6%) ולכן נחליט להסיר את המשתנה.

משתנים קטגוריאליים:



איור 28 Y~regime-Scatterplot



איור 27 Y~religion-Scatterplot

X_{11} : Regime

בבדיקת הנתונים הסטטיסטיים התגלה כי ה- $R^2_{adj} = 0.06$ וה-Pvalue שלו נמוך (3.3%). אמנם ה- R^2_{adj} יחסית נמוך, אך נראה כי יש קשר בין המשתנים ולכן נבחר לא להסיר משתנה זה.

X₁₀: Religion

ניתן לראות בגרף הפיזור כי קיימות דתות עם תצפיות מעטות מאוד, מה שמקשה על הסקה, בנוסף בדתות בהן יש מספיק תצפיות על מנת להסיק, טווח הערכים רחב ומפוזר לאורך כלל הערכים שהמשתנה המוסבר מקבל, לכן לא הצלחנו לראות קשר לינארי. בנוסף ה- $R^2_{adj}=0.017$ מה שמראה כי המשתנה מסביר מעט על המשתנה המוסבר, בנוסף ה-Pvalue גבוה (20%) ולכן נבחר להסיר משתנה זה.

לסיכום - המשתנים שהסרנו הם: Hrs of work, Divorce rates, Religion

10.2 התאמת משתנים:

אנו מציגים את היישום בקוד [בנספח ח'](#).

X₁₁: Regime איחוד קטגוריות במשתנה הקטגורי

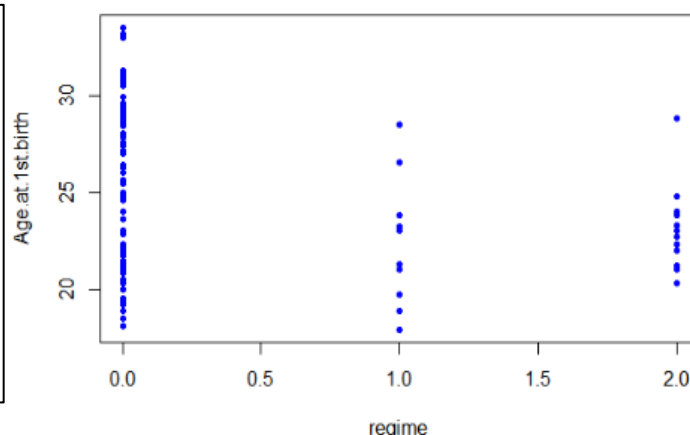
עבור משתנה זה קיבלנו $R^2_{adj}=0.06$, ונרצה לבחון האם איחוד הקטגוריות יוביל לעלייה בערך זה. ראשית מהתבוננות על גרף הפיזור ניתן לראות כי קיימות שתי קטגוריות של משטרים (כבוש (3) וארעי (8)) אשר בעלות תצפיות מינוריות (תצפית אחת או שתיים) ולכן לא ניתן להסיק מהן כראוי ועל כן נבחר לנפותן. בהסתכלות על הקטגוריות הנוספות מרבית התצפיות בבסיס הנתונים שלנו בעלות משטרי דמוקרטיה (1) ומפלגה דומיננטית (2). ניתן לראות בגרף הפיזור כי התצפיות משני משתנים אלה מתפרשות לאורך טווח ערכים רחב וזהה. מנגד קיימות קטגוריות (ממשל צבאי (4), מונרכיה (5) חד-מפלגתי (6)) המקבלות ערכים בטווח יחסית זהה של המשתנה המוסבר (גילאים 20-25) ועל כן נבחר לשלבן לקטגוריה אחת. בנוסף קיימת קטגוריית דיקטטורה (7) בה מתקבלים גם ערכים נמוכים בשונה מהאיחוד השני, אך לא מתקבלים כלל ערכים גבוהים בשונה מהקטגוריות הראשונות שאיחדנו ולכן לא נצרפה לאף אחד מהאיחודים.

```
Call:
lm(formula = Age.at.1st.birth ~ (regime), data = unionDataSet,
    x = TRUE, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-7.1262 -3.3262  0.3738  3.3738  8.2638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.2262     0.3988   63.260 < 2e-16 ***
regime       -1.3162     0.5016   -2.624  0.00979 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.975 on 123 degrees of freedom
Multiple R-squared:  0.05301, Adjusted R-squared:  0.04531
F-statistic: 6.885 on 1 and 123 DF, p-value: 0.009792
```



איור 30 - Summary(Y~regime)

איור 29 - Scatterplot Y~regime

בבדיקת הנתונים הסטטיסטים ניתן לראות כי אמנם ה- R^2_{adj} ירד ב-0.015 אך לעומתו ה-Pvalue צנח מ-3.3% ל-0.9% ולכן נבחר כן ליישם את איחוד הקטגוריות.

X₁: Rate of happiness הפיכת משתנה רציף למשתנה קטגורי

לאחר מעבר על טבלת הקורלציות עבור המשתנים שבחרנו להשאיר במודל בסעיף הקודם, הבחנו כי משתנה "מדד האושר" הנו בעל הקורלציה הנמוכה ביותר למשתנה המוסבר, ועל כן נבדוק האם הפיכתו לקטגוריאלי תשפר את המודל, ותוביל להסברה טובה יותר של מדד אושר על המשתנה המוסבר. בבדיקת הנתונים הסטטיסטיים ניתן

לראות כי ה- R^2_{adj} ירד ואף ה-Pvalue עלה, ועל כן נבחר שלא לשנות את המודל ולהשאירו כפי שהיה.

```
lm(formula = AgeAtFirstBirth ~ Rate.of.happiness, data = datasetNew,
   x = TRUE, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-7.4295 -2.9051  0.0949  2.3705 10.5849

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    22.9051     0.4571  50.113  <2e-16 ***
Rate.of.happiness  3.7244     0.6490   5.739   7e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.628 on 123 degrees of freedom
Multiple R-squared:  0.2112,    Adjusted R-squared:  0.2048
F-statistic: 32.93 on 1 and 123 DF,  p-value: 7e-08
```

איור 31 - Summary(Y~Rate.of.happiness)

10.3

הגדרת משתני דמה:

אנו מציגים את היישום בקוד [בנספח ח'](#).

ניתן לראות את תרשים הפיזור של המשתנה "משטר" בעמוד למעלה.

המשתנה הקטגוריאל שנוצר הוא המשטר. נגדיר את המשתנה הבינארי המתאים כדי שנוכל לאמוד את תרומת משתנה זה למודל הרגרסיה, כך שמשתנה הבסיס הוא "דמוקרטיה וגם מפלגה דומיננטית":

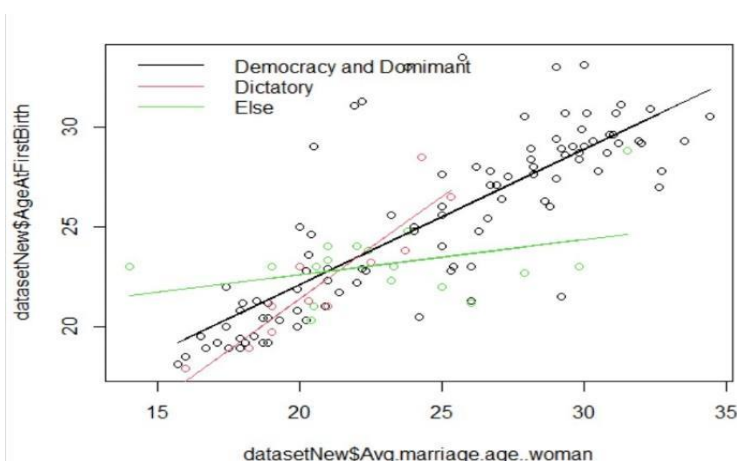
$$rgmUni = \begin{cases} 1, & \text{Unioned Categories} \\ 0, & \text{Else} \end{cases}$$

$$rgmDict = \begin{cases} 1, & \text{Personal Dictatorship} \\ 0, & \text{Else} \end{cases}$$

10.4

הוספת משתני אינטראקציה:

כל ההשוואות בוצעו ביחס למשתנה הקטגוריאל "משטר".

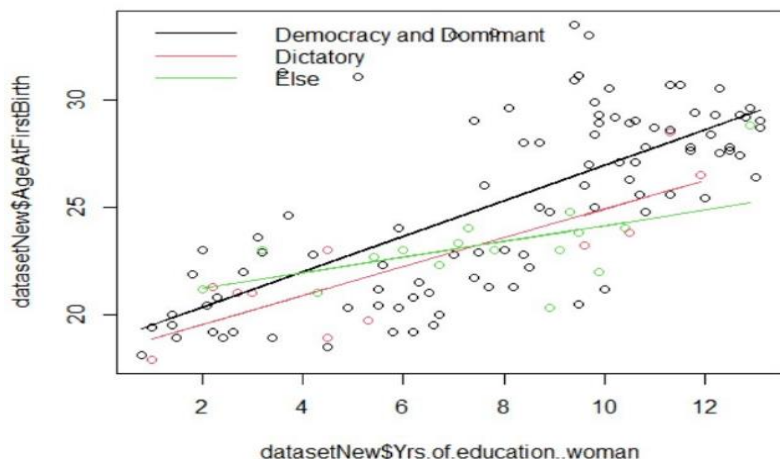


איור 32 - Scatterplot marriage~Y~regime

השוואה א' - השפעת גיל הנישואים הממוצע

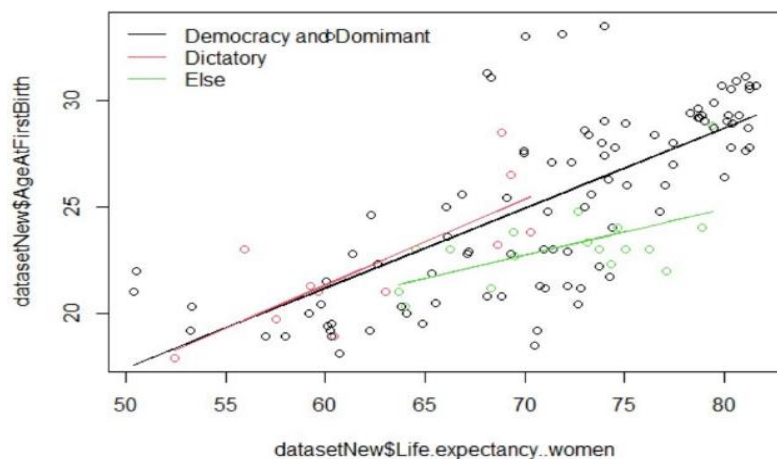
לנשים (X_7) על המשתנה המוסבר, כתלות במשטר (X_{11}).

ניתן לראות שההבדל בסוג המשטר גורם לשינוי במגמת הגרף, במיוחד בין גרף הבסיס (משטר דמוקרטי / מפלגה דומיננטית) לבין הגרף שמייצג את שאר המשטרים (Else), ולכן נחליט להכניס משתנה אינטראקציה זה למודל.



איור 33 - Scatterplot education~regime~Y

השוואה ב' - השפעת שנות השכלה (X_3) על המשתנה המוסבר, כתלות במשטר (X_{11}). אמנם ניתן לראות הבדל מבחינת השיפוע בין גרף הבסיס (משטר דמוקרטי/מפלגה דומיננטית) לבין הגרף שמייצג את שאר המשטרים (Else), אך עם זאת נשים לב שה-PVALUE של המשתנה שמייצג את התוספת השולית של משטר דיקטטורי לשיפוע גרף הבסיס גבוה מאוד, מה שמעיד על כך שאין השפעה על שיפוע גרף הבסיס. לכן החלטנו שלא להכניס את המשתנה הזה למודל.

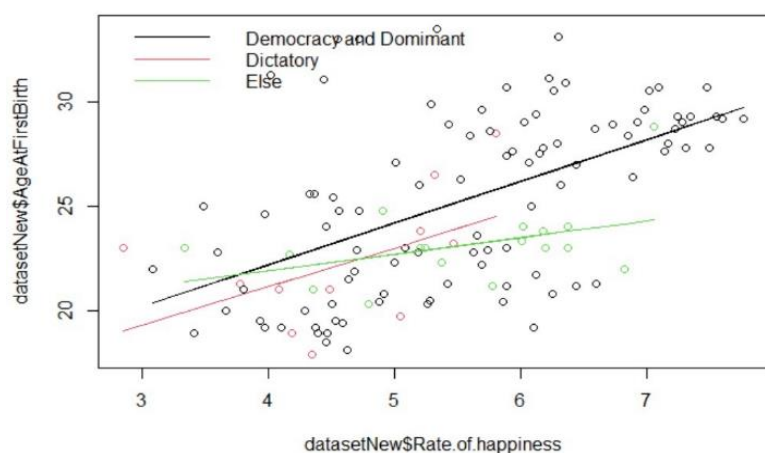


איור 34 - Scatterplot Life.expectancy~regime~Y

השוואה ג' - השפעת תוחלת החיים (X_5) על המשתנה המוסבר, כתלות במשטר (X_{11}). ניתן לראות שאין הבדל גדול מבחינת השיפוע בין גרף הבסיס (משטר דמוקרטי/מפלגה דומיננטית) לבין שאר הגרפים, במיוחד בגרף שמייצג משטר דיקטטורי שכמעט מתלכד עם גרף הבסיס, ולכן נחליט לא להכניס את המשתנה למודל.

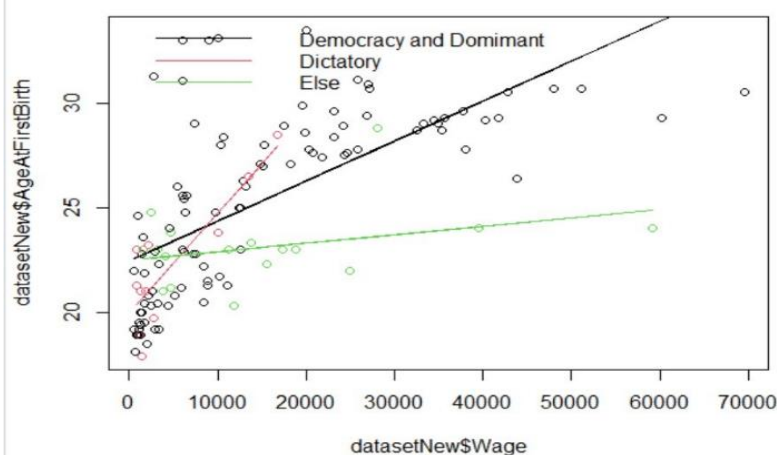
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	18.69489	0.73222	25.532	<2e-16 ***
Yrs.of.education..woman	0.82581	0.08358	9.880	<2e-16 ***
regimeFactor2	-0.49231	1.79168	-0.275	0.784
regimeFactor3	1.79820	2.04693	0.878	0.381
Yrs.of.education..woman:regimeFactor2	-0.15556	0.24348	-0.639	0.524
Yrs.of.education..woman:regimeFactor3	-0.46144	0.25953	-1.778	0.078 .

איור 35 - Summary for previous scatterplot



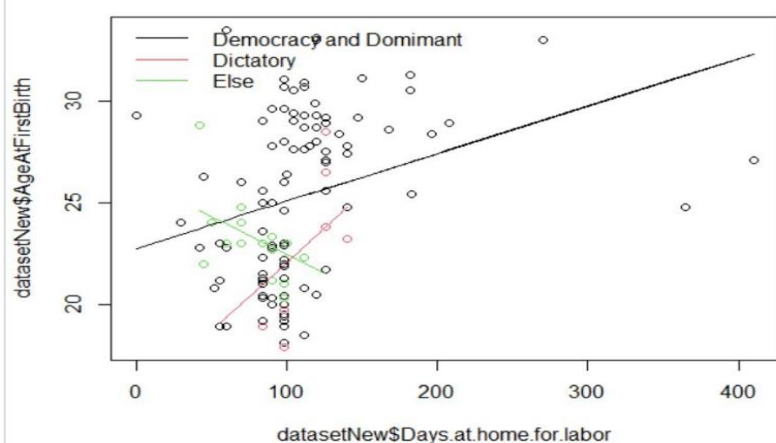
איור 36 - Scatterplot happiness~regime~Y

השוואה ד' - השפעת דרגת האושר (X_{11}) על המשתנה המוסבר, כתלות במשטר (X_{11}). ניתן לראות שאין הבדל גדול מבחינת השיפוע בין גרף הבסיס (משטר דמוקרטי/מפלגה דומיננטית) לבין שאר הגרפים, ולכן **נחליט לא להכניס את המשתנה למודל**.



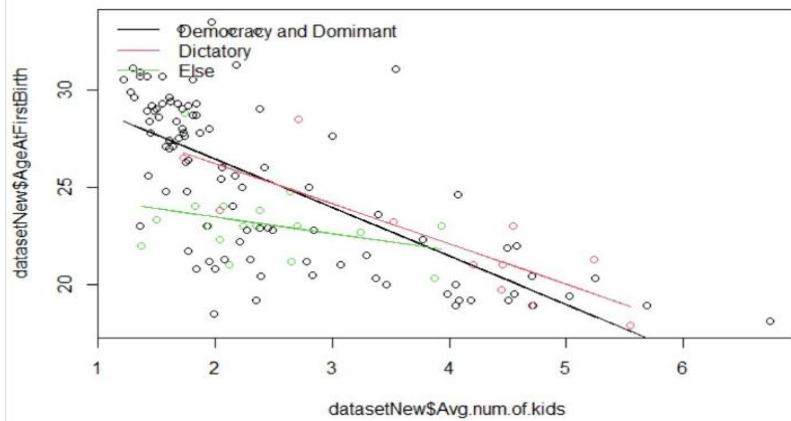
איור 37 - Scatterplot wage~regime~Y

השוואה ה' - השפעת השכר (X_9) על המשתנה המוסבר, כתלות במשטר (X_{11}). ניתן לראות שקיים הבדל גדול מבחינת השיפוע בין כל הגרפים, ולכן **נחליט להכניס את המשתנה למודל**.



איור 38 - Scatterplot labor~regime~Y

השוואה ו' - השפעת חופשת הלידה (X_1) על המשתנה המוסבר, כתלות במשטר (X_{11}). ניתן לראות שקיים הבדל גדול מבחינת השיפוע בין כל הגרפים, ולכן **נחליט להכניס את המשתנה למודל**.



איור 39 - Scatterplot kids~regime~Y

השוואה ז' - השפעת מס' הילדים (X_8) על המשתנה המוסבר, כתלות במשטר (X_{11}).

ניתן לראות שאין הבדל גדול מבחינת השיפוע בין גרף הבסיס (משטר דמוקרטי/מפלגה דומיננטית) לבין שאר הגרפים, במיוחד בגרף שמייצג משטר דיקטטורי שכמעט מתלכד עם גרף הבסיס. כמו כן כלל הגרפים יורדים. לכן נחליט לא להכניס את המשתנה למודל.

המודל הנוכחי:

$$\begin{aligned} \hat{y} = & \hat{\beta}_1 + \hat{\beta}_2 * rgmUni + \hat{\beta}_3 * rgmDict \\ & + \hat{\beta}_4 * x_1 + \hat{\beta}_5 * x_1 * rgmUni + \hat{\beta}_6 * x_1 * rgmDict + \hat{\beta}_7 * x_2 + \hat{\beta}_8 * x_3 + \hat{\beta}_9 * x_5 \\ & + \hat{\beta}_{10} * x_7 + \hat{\beta}_{11} * x_7 * rgmUni + \hat{\beta}_{12} * x_7 * rgmDict + \hat{\beta}_{13} * x_8 \\ & + \hat{\beta}_{14} * x_9 + \hat{\beta}_{15} * x_9 * rgmUni + \hat{\beta}_{16} * x_9 * rgmDict \end{aligned}$$

11. התאמת המודל ובדיקת הנחות המודל:

בנספח ו' אנו מציגים את תוצאות האלגוריתמים.

11.1 בחירת משתני המודל:

כדי להגיע למודל הסופי ניעזר באלגוריתמים שונים לצורך החלטה האם להוסיף או להשמיט משתנים מהמודל. נבצע שיטות אלו על המודל הנוכחי שלנו, כאשר בכל שיטה נשווה בין האפשרויות בעזרת המדד AIC ע"מ להחליט אילו מבין המודלים הנו האידאלי.

גרסיה לפני:

ראשית, נריץ את המודל ללא משתנים ובכל אינטראקציה יתווסף עד משתנה אחד למודל, כאשר המשתנה בעל מדד ה-AIC הנמוך ביותר הוא זה שייבחר. לאחר סיום ההרצה, התקבלו המשתנים הבאים:

גיל הנישואין הממוצע לנשים, שני סוגי משתני האינטראקציה "משטר" שמוכפלים במשתנה שמייצג את גיל הנישואין הממוצע לנשים, מס' ילדים ממוצע, ימי חופשת לידה בתשלום לנשים ושכר.

מדד AIC: 212.82

רגרסיה לאחור:

ראשית, נריץ את המודל שמכיל את כל המשתנים ובכל איטרציה נשמיט מהמודל עד משתנה אחד, כאשר המשתנה בעל מדד ה-AIC הגבוה ביותר הוא זה שיושמט. לאחר סיום ההרצה, התקבלו המשתנים הבאים:

תוחלת חיים, ימי חופשת לידה בתשלום לנשים, גיל הנישואין הממוצע לנשים, שני סוגי משתני האינטראקציה "משטר" שמוכפלים במשתנה שמייצג את גיל הנישואין הממוצע לנשים/

מדד AIC: 211.76

רגרסיה בצעדים:

ראשית, נריץ את המודל שמכיל את כל המשתנים ובכל איטרציה נשמיט או נוסיף משתנה יחיד, לאחר שנבחן כיצד נוכל למזער את מדד ה-AIC באופן המיטבי. לאחר סיום ההרצה, התקבלו המשתנים הבאים:

ימי חופשת לידה בתשלום לנשים, תוחלת חיים, גיל הנישואין הממוצע לנשים, שני סוגי משתני האינטראקציה "משטר" שמוכפלים במשתנה שמייצג את גיל הנישואין הממוצע לנשים.

מדד AIC: 211.76

המודל הנבחר:

בחרנו את מודל Stepwise שיקבע עבורנו את מודל הרגרסיה מכיוון שהוא בעל מדד ה-AIC הנמוך ביותר. ערך המדד באלגוריתם זה זהה לערך המדד שמתקבל באלגוריתם רגרסיה לאחור, אך למדנו שרגרסיה בצעדים היא יותר מדויקת ולכן בחרנו בה.

לאחר סיום ההרצה, התקבלו המשתנים הבאים:
ממוצע ימי חופשת לידה בתשלום לנשים (X_1), תוחלת חיים לנשים (X_5), גיל נישואין ממוצע לנשים (X_7), 2 משתני הדמה של משטר ($rgmDict, rgmUni$), משתנה האינטראקציה ($rgmDict * X_7, rgmUni * X_7$).

$$\hat{y} = 3.46758 + (-4.977042 * rgmUni) + (7.642933 * rgmDict) + (0.008311 * X_1) \\ + (0.108996 * X_5) + (0.533411 * X_7) + (0.239760 * rgmUni * X_7) \\ + (-0.380903 * rgmDict)$$

```

Residuals:
    Min       1Q   Median       3Q      Max
-4.7854 -1.3319 -0.1523  0.8088  7.7483

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.467580   2.081829   1.666  0.09846 .
Days.at.home.for.labor 0.008311   0.004214   1.972  0.05097 .
regimeFactor2   -4.977042   5.471461  -0.910  0.36488
regimeFactor3    7.642933   3.495070   2.187  0.03075 *
Life.expectancy..women 0.108996   0.040731   2.676  0.00852 **
Avg.marriage.age..woman 0.533411   0.066084   8.072 6.88e-13 ***
regimeFactor2:Avg.marriage.age..woman 0.239760   0.257514   0.931  0.35374
regimeFactor3:Avg.marriage.age..woman -0.380903   0.147478  -2.583  0.01103 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.262 on 117 degrees of freedom
Multiple R-squared:  0.7084,    Adjusted R-squared:  0.6909
F-statistic: 40.6 on 7 and 117 DF,  p-value: < 2.2e-16

```

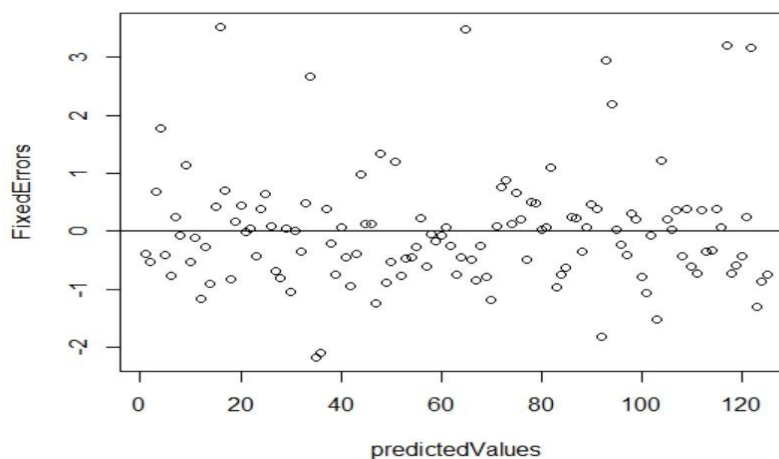
איור 40 - Summary of stepwise

11.2 בדיקת הנחות המודל:

(בנספח ז' קטעי הקוד הרלוונטים)

בדיקת הנחת הלינאריות:

על מנת לבדוק את הנחת הלינאריות ראשית מצאנו מה הערכים הנצפים של המודל. לאחר מכן בדקנו את השגיאות של התצפיות שלנו ותיקנו אותם. בנינו ע"ס נתונים אלה תרשים פיזור של הערכים הנצפים אל מול השגיאות; ניתן לראות בתרשים כי רוב הנתונים פזורים באופן יחסית אחיד סביב קו ה-0, ולכן נסיק שהנחת הלינאריות מתקיימת.



איור 41 - Scatterplot Yhat~FixedErrors

בדיקת הנחת שוויון שוננויות:

לפי הנחת שוויון השוננויות לכל ε_i מתקיים $v(\varepsilon_i) = \sigma^2$.
שוויון שוננויות מתקיים כאשר יש פיזור אחיד לאורך הציר, וניתן לראות בבחינת תרשים פיזור
השאריות כי הפיזור יחסית זהה סביב ה-0 ולא משתנה לאורך ציר ה-X, קיימות גם שאריות עבורן
לא מתקיים הפיזור האחיד אך הן מעטות ולכן ניתן לשער שההנחה מתקיימת. נבחן הנחה זו ע"י
מבחן F: תחילה בוצעה חלוקת ערכי הערך המוסבר לשליש עליון ושליש תחתון, לאחר מכן, על
ערכים אלה התשמנו במבחן F ב-R בו חושב היחס בין השוננויות. תוצאות המבחן הציגו Pvalue
גבוה מ-5% ועל כן לא ניתן לדחות את השערת ה-0, כלומר מתקיים שוויון שוננויות.

```
F test to compare two variances

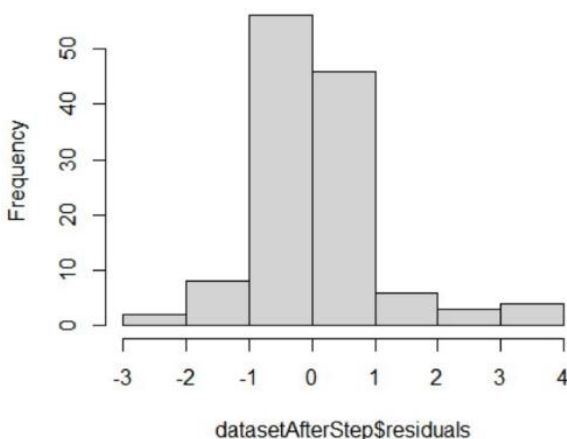
data:  thirdData and twothirdData
F = 0.54151, num df = 41, denom df = 42, p-value = 0.05175
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2925238 1.0047513
sample estimates:
ratio of variances
 0.5415109
```

איור 42 - Ftest

בדיקת הנחת הנורמליות:

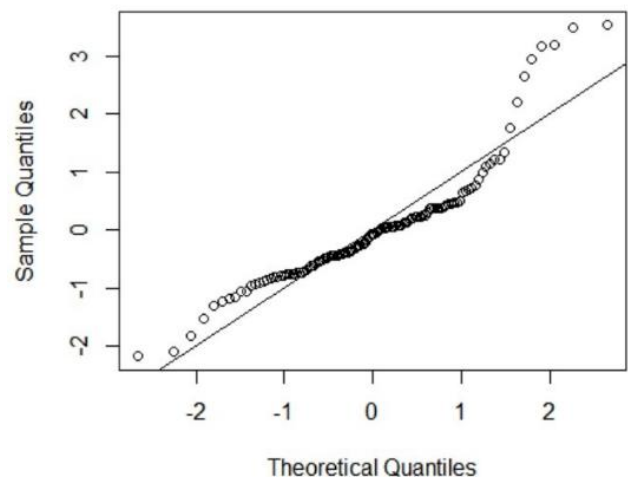
על מנת לבדוק את הנחת הנורמאליות ראשית מצאנו מה הערכים הנצפים של המודל. לאחר מכן
בדקנו את השגיאות של התצפיות שלנו ותיקנו אותם. בנינו QQplot ותרשים היסטוגרמה כדי
לבחון באופן תיאורי את הנחת הנורמאליות, וראינו שהגרפים לא מציגים נתונים שתואמים
להתפלגות הנורמאלית. לבסוף, בחנו את הנחת הנורמאליות ע"י מבחן KS ותוצאת המבחן העידה
על כך שהשגיאות אינן מתפלגות נורמאלית בר"מ 5%.

Histogram of datasetAfterStep\$residuals



איור 43 - Histogram of residuals

Normal Q-Q Plot



איור 44 - QQplot of errors

One-sample Kolmogorov-Smirnov test

```
data:  datasetAfterStep$residualsFix
D = 0.14966, p-value = 0.007398
alternative hypothesis: two-sided
```

KS test - 45 איור

11.3 דוגמה לשימוש במודל הנבחר:

(בנספח ט' קטעי הקוד הרלוונטי)

בעזרת המודל בחנו את הגיל הממוצע ללידה ראשונה באיחוד אמירויות, מכיוון שכעת עם הקשר והיחסים החדשים בין מדינתו לאמירויות, רבים התחילו להתעניין במדינה, וייתכן וזוגות ירצו לעבור לגור שם. במעבר כזה יתעניינו לגבי הנורמות במדינה הנוגעת להקמת משפחה.

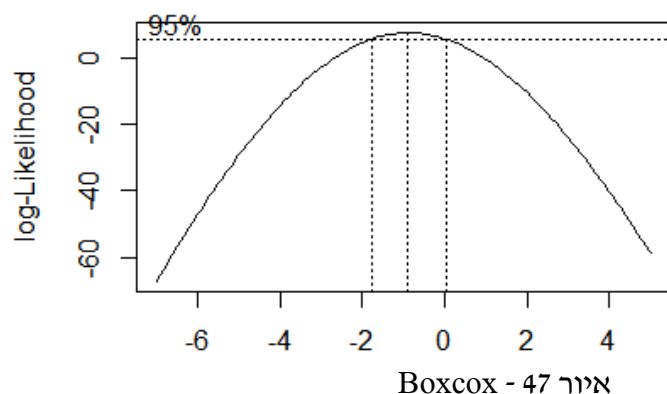
חיפשנו בטבלה המקורית טרם הסרת שמות המדינות את השורה בה נמצאת המדינה, והוצאנו את המידע מהטבלה לאחר כלל הצעדים לשיפור המודל שנערכו עד כה. התקבלה תחזית של גיל 23.69 ממוצע ללידה, בעוד הערך הנמצא בבסיס הנתונים הינו גיל 22. ניתן לראות גם כי התחזית יחסית קרובה לערך הקיים, ועל כן נראה כי נוכל לענות לזוגות אלה עם תחזית קרובה.

```
> datasetTotal$Country.Name[rowArab]
[1] "United Arab Emirates"
> print(datasetAfterStep[rowArab:rowArab,1:length(datasetAfterStep)-1]) # dataOf United Arab Emirates
AgeAtFirstBirth Days.at.home.for.labor regime Life.expectancy..women Avg.marriage.age..woman predicted
6 22 45 3 77.06 25 23.69645
```

איור Data for UAE-46

12. שיפור המודל:

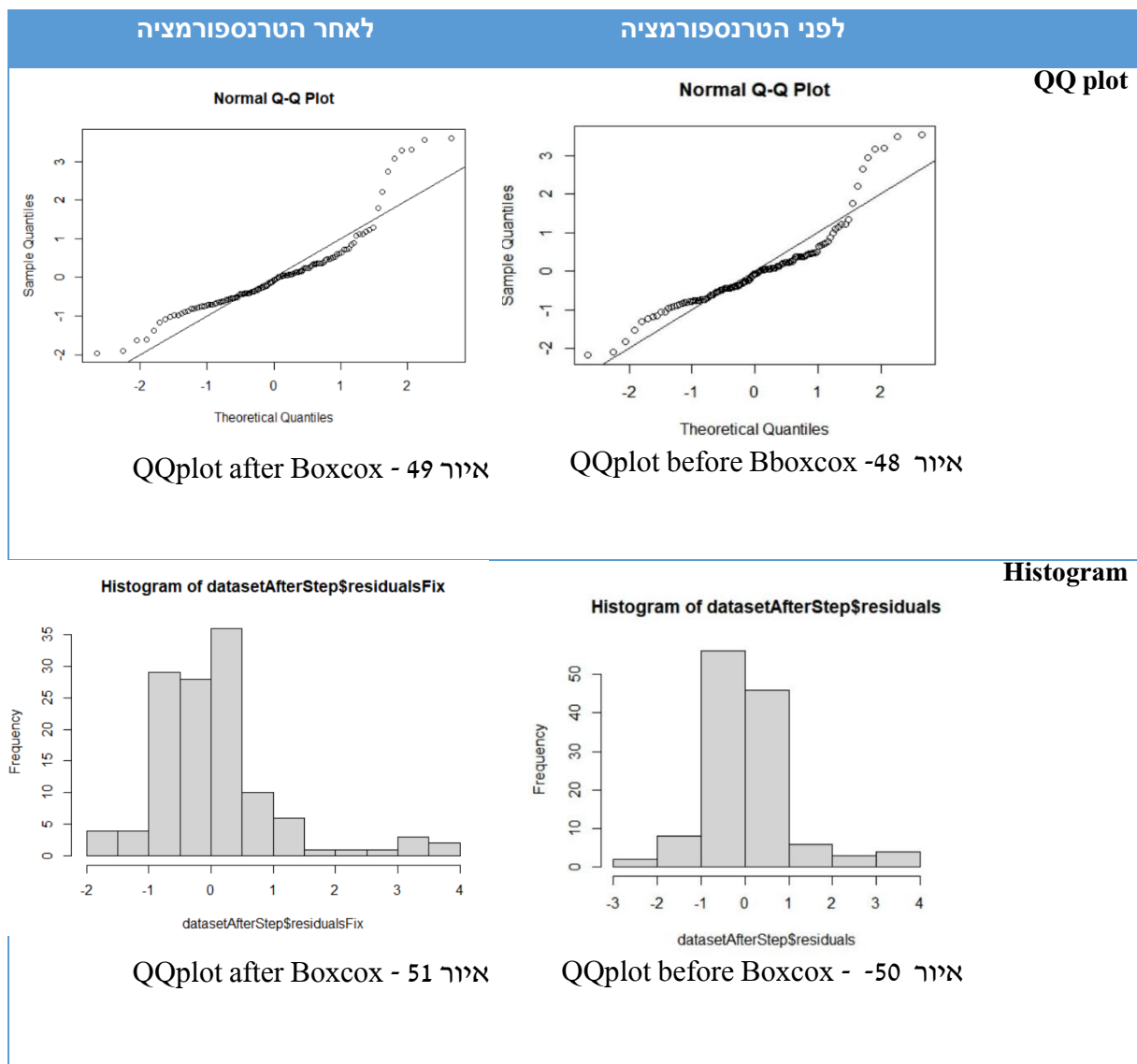
(בנספח ט' קטעי הקוד הרלוונטיים)



לאחר בחינת הנחות המודל הסופי, נראה כי המודל אינו עומד בהנחת נורמאליות השגיאות ולכן נרצה לבצע cox-box על המודל, באמצעות נבחן איזה טרנספורמציה תתאים ביותר למשתנה המוסבר על מנת לאפשר למודל לעמוד בכלל ההנחות.

ניתן לראות כי הערך $\lambda = 0$ קיים ברווח הסמך, ולכן נבחר להשתמש בו.

נציג בטבלת השוואה את גרף ה QQplot וגרף ההיסטוגרמה לפני הטרנספורמציה ואחריה.



בהסתכלות על QQ plot המשקף את הנחת הנורמאליות של השגיאות המתוקנות, ובהסתכלות על גרף ההיסטוגרמה, ניתן לראות שיפור מזערי בהנחת הנורמאליות. נמשיך למבחן הסטטיסטי KS לבדיקת ההנחה:

One-sample Kolmogorov-Smirnov test

```
data: datasetAfterStep$residualsFix
D = 0.12652, p-value = 0.03655
alternative hypothesis: two-sided
```

איור 52 - New KS test

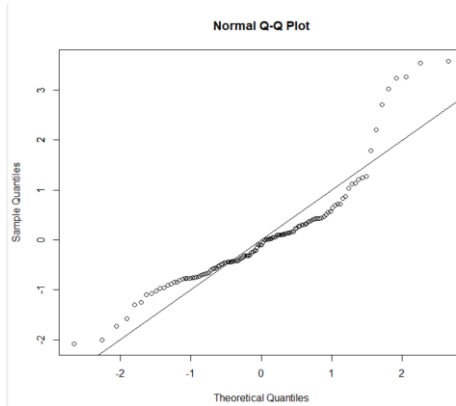
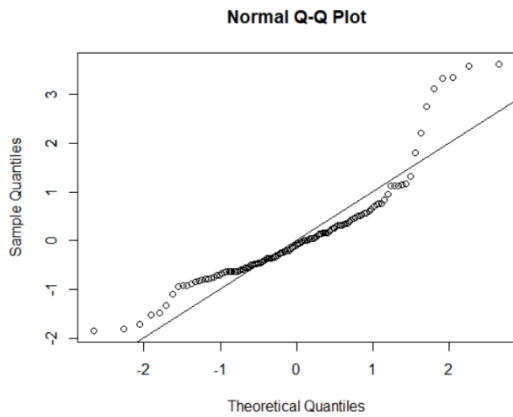
ניתן לראות מפלט המבחן כי ניכר שיפור משמעותי בערכו של Pvalue (ערך של 0.03655 לעומת ערך של 0.0073). אולם עדיין איננו עומדים בהנחת הנורמאליות, שכן אנו עדיין נדחה את השערת האפס בר"מ 5% שמדובר בהתפלגות נורמאלית.

כעת ננסה לבצע טרנספורמציות מוכרות שנלמדו בכיתה, בשאיפה שאחת מהן תאפשר לנו לדחות את השערת האפס ולעמוד בכלל הנחות המודל.

$$f(y) = y^{-0.5}$$

$$f(y) = \sqrt{y}$$

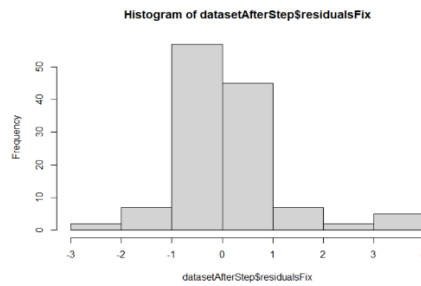
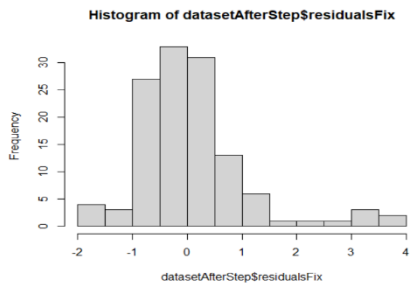
QQ plot



QQplot for 2nd trans -54 איור

QQplot for 1st trans -53 איור

Histogram



Histogram for 2nd trans -56 איור

Histogram for 1st trans -55 איור

KS test

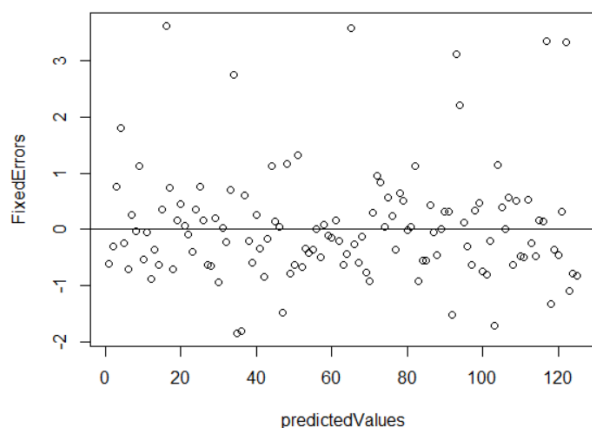
One-sample Kolmogorov-Smirnov test

data: datasetAfterStep\$residualsFix
D = 0.1189, p-value = 0.05836
alternative hypothesis: two-sided

One-sample Kolmogorov-Smirnov test

data: datasetAfterStep\$residualsFix
D = 0.13762, p-value = 0.01757
alternative hypothesis: two-sided

עם בדיקת הטנספורמציה $f(y) = y^{-0.5}$ נוכחנו לגלות כי אנו עומדים בהנחת הנורמאליות של השגיאות המתוקנות, שכן לפי מבחן KS לא נדחה את השערת האפס ונגיד כי השגיאות מתפלגות נורמאלית. כעת נותר לבדוק את הנחת הלינאריות (הנחת שיוויון השנויות נשארת כפי שהייתה, שכן וקטור המשתנה המוסבר אינו משתנה), על מנת לבחור סופית טרנספורמציה זו.

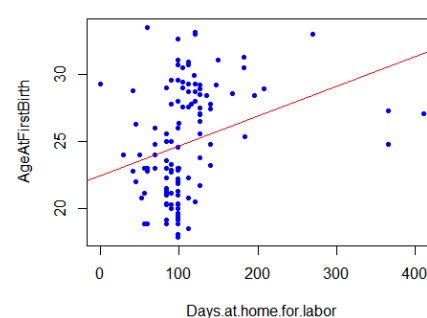
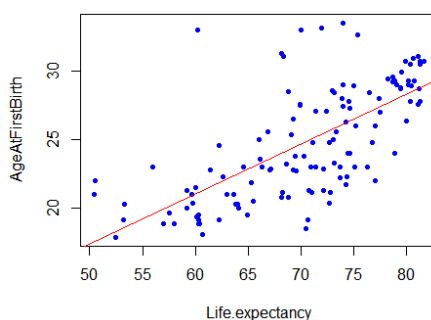
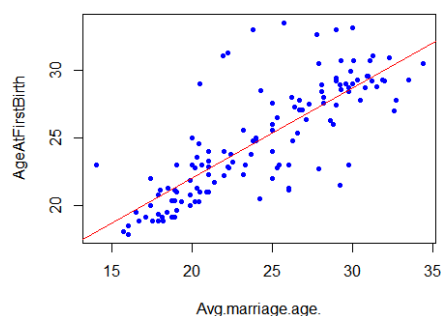


בנינו גרף פיזור של השגיאות המתוקננות, וניתן לראות כי רוב הנתונים פזורים באופן יחסית אחיד סביב קו ה-0, ולכן נסיק שהנחת הלינאריות מתקיימת.

לכן אנו עומדים בהנחות המודל

איור 57 - Final scatterplot Yhat~fixed errors

בהמשך למטרתנו לשיפור המודל, בחנו את האפשרות לבצע טרנספורמציה על המשתנים המסבירים שנותרו במודל. התבוננו בגרפי הפיזור של המשתנים עם המוסבר, ולא נראתה התנהגות המזכירה פונקציה מוכרת, ועל כן נבחר שלא לבצע טרנספורמציה זו.



לכן המודל הסופי הינו:

$$\hat{y}^{-0.5} = 3.46758 + (-4.977042 * rgmUni) + (7.642933 * rgmDict) + (0.008311 * X1) + (0.108996 * X5) + (0.533411 * X7) + (0.239760 * rgmUni * X7) + (-0.380903 * rgmDict)$$

13. נספחים

13.1 נספח א' - תרשים קורלציה

	AgeAtFirstBirth	Days.at.home.for.labor	Rate.of.happiness	Yrs.of.education..woman	Divorce.rates	Life.expectancy..women	Hrs.of.work	Avg.marriage.age..woman	Avg.num.of.kids	Wage
AgeAtFirstBirth	1.0000000	0.2961647	0.5254850	0.6643153	0.1638211	0.6871282	-0.2413891	0.7966313	-0.6589406	0.6102858
Days.at.home.for.labor	0.2961647	1.0000000	0.0381499	0.2165497	-0.0992616	0.1572523	-0.2685769	0.2253880	-0.2554800	0.0418313
Rate.of.happiness	0.5254850	0.0381499	1.0000000	0.6781951	0.1959441	0.7952229	-0.2220739	0.6265957	-0.6322696	0.7481269
Yrs.of.education..woman	0.6643153	0.2165497	0.6781951	1.0000000	0.1414127	0.7217345	-0.3217178	0.7148287	-0.7429480	0.7007072
Divorce.rates	0.1638211	-0.0992616	0.1959441	0.1414127	1.0000000	0.1305055	-0.0457551	0.2030871	-0.1263886	0.2201289
Life.expectancy..women	0.6871282	0.1572523	0.7952229	0.7217345	0.1305055	1.0000000	-0.1320051	0.7187997	-0.8190851	0.7237323
Hrs.of.work	-0.2413891	-0.2685769	-0.2220739	-0.3217178	-0.0457551	-0.1320051	1.0000000	-0.3308451	0.0664905	-0.2305328
Avg.marriage.age..woman	0.7966313	0.2253880	0.6265957	0.7148287	0.2030871	0.7187997	-0.3308451	1.0000000	-0.6886983	0.6604981
Avg.num.of.kids	-0.6589406	-0.2554800	-0.6322696	-0.7429480	-0.1263886	-0.8190851	0.0664905	-0.6886983	1.0000000	-0.5778841
Wage	0.6102858	0.0418313	0.7481269	0.7007072	0.2201289	0.7237323	-0.2305328	0.6604981	-0.5778841	1.0000000

13.2 נספח ב' - דוגמה לקטע הקוד של ניתוח תיאורי של המשתנים

```

answer <- function(vector,string){ #to display data comfortably
  MEAN <- mean(vector)
  MEDIAN <- median(vector)
  SD <- sd(vector)
  QUANTILE <- quantile(vector,probs=c(0.25,0.5,0.75))
  SKEWNESS <- skewness(vector)
  print(string)
  paste("mean=",MEAN," median=",MEDIAN," sd=",SD," quantile=", QUANTILE," skewness=",SKEWNESS)
}

print(answer(datawithoutCat$AgeAtFirstBirth,"AgeAtFirstBirth"))
print(answer(datawithoutCat$Days.at.home.for.labor,"Days.at.home.for.labor"))
print(answer(datawithoutCat$Rate.of.happiness,"Rate.of.happiness"))
print(answer(datawithoutCat$Yrs.of.education..woman,"Yrs.of.education..woman"))
print(answer(datawithoutCat$Divorce.rates,"Divorce.rates"))
print(answer(datawithoutCat$Life.expectancy..women,"expectancy..women"))
print(answer(datawithoutCat$Hrs.of.work,"Hrs.of.work"))
print(answer(datawithoutCat$Avg.marriage.age..woman,"Avg.marriage.age..woman"))
print(answer(datawithoutCat$Avg.num.of.kids,"Avg.num.of.kids"))
print(answer(datawithoutCat$Wage,"Wage"))

```

```

categoricalDataset <- subset(dataset,select=c("AgeAtFirstBirth","regime"))

yAtRegime1 <- sqldf("select AgeAtFirstBirth from categoricalDataset where regime='Democracy'")
print(answer(yAtRegime1$AgeAtFirstBirth,"AgeAtFirstBirth"))

yAtRegime2 <- sqldf("select AgeAtFirstBirth from categoricalDataset where regime='Dominant Party'")
print(answer(yAtRegime2$AgeAtFirstBirth,"AgeAtFirstBirth"))

yAtRegime3 <- sqldf("select AgeAtFirstBirth from categoricalDataset where regime='Foreign/Occupied'")
print(answer(yAtRegime3$AgeAtFirstBirth,"AgeAtFirstBirth"))

yAtRegime4 <- sqldf("select AgeAtFirstBirth from categoricalDataset where regime='Military'")
print(answer(yAtRegime4$AgeAtFirstBirth,"AgeAtFirstBirth"))

```

13.3 נספח ג'- פלט הטבלאות החד מימדיות

```
> #1D tables
> x2.9 <- binFreqTable(dataset$Rate.of.happiness,seq(2,8,by=1))
> hist(dataset$Rate.of.happiness); axis(1,at=seq(2,8,1))
> x2.9
  range frequency
1 2 - 3         1
2 3 - 4        11
3 4 - 5        34
4 5 - 6        35
5 6 - 7        32
6 7 - 8        15

> xx2.9 <- binFreqTable(dataset$Avg.num.of.kids,seq(0,7,by=1))
> hist(dataset$Avg.num.of.kids); axis(1,at=seq(0,7,1))
> xx2.9
  range frequency
1 0 - 1         0
2 1 - 2        58
3 2 - 3        36
4 3 - 4        12
5 4 - 5        16
6 5 - 6         5
7 6 - 7         1
```

13.4 נספח ד'- פלט הטבלאות הדו מימדיות

```
> ANS <- matrix( 0,nrow = 7, ncol = 7)
> x=20
> y=0
> for(n in 1:7){
+   for(k in 1:7){
+     for(i in 1:128){
+       if((dataWithoutCat[i,7]>=x && dataWithoutCat[i,7]<(5+x)) && (dataWithoutCat[i,10]>=y && dataWithoutCat[i,10]<(10000+y))){
+         ANS[n,k]=ANS[n,k]+1
+       }
+     }
+     y=y+10000
+   }
+   y=0
+   x=x+5
+ }
> ANS
  [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  0   0   0   0   0   0   0
[2,]  0   0   0   0   0   0   0
[3,]  0   0   0   0   0   0   0
[4,]  2   0   1   3   2   0   1
[5,] 31  13  13   4   2   1   1
[6,] 30  13   3   2   0   1   0
[7,]  4   0   0   0   1   0   0
```

```

> ANS2 <- matrix( 0,nrow = 7, ncol = 7)
> x=14
> y=0
> for(n in 1:7){
+   for(k in 1:7){
+     for(i in 1:128){
+       if((dataWithoutCat[i,8]>=x && dataWithoutCat[i,8]<(3+x)) && (dataWithoutCat[i,9]>=y && dataWithoutCat[i,9]<(1+y))){
+         ANS2[n,k]=ANS2[n,k]+1
+       }
+     }
+     y=y+1
+   }
+   y=0
+   x=x+3
+ }
> ANS2
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  0    2    0    0    1    2    1
[2,]  0    2    6    2   12    2    0
[3,]  0    3   11    8    3    1    0
[4,]  0    6   12    1    0    0    0
[5,]  0   15    5    1    0    0    0
[6,]  0   23    2    1    0    0    0
[7,]  0    6    0    0    0    0    0

```

13.5 נספח ה'- Summaries של משתנים שמועמדים להסרה

מס' שעות העבודה:

```

Call:
lm(formula = Age.at.1st.birth ~ Hrs.of.work, data = dataset,
    x = TRUE, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-8.0096 -2.9318 -0.1819  2.8119  9.7653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.37743   3.78034   9.358 3.99e-16 ***
Hrs.of.work -0.24277   0.08695  -2.792 0.00605 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.998 on 126 degrees of freedom
Multiple R-squared:  0.05827, Adjusted R-squared:  0.05079
F-statistic: 7.796 on 1 and 126 DF, p-value: 0.006052

```

מס' ימי חופשת הלידה:

```

Call:
lm(formula = Age.at.1st.birth ~ Days.at.home.for.labor, data = dataset,
    x = TRUE, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-6.7465 -3.2864 -0.7272  3.3362  9.6840

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.478967   0.769567  29.210 < 2e-16 ***
Days.at.home.for.labor 0.022118   0.006355   3.481 0.000688 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.935 on 126 degrees of freedom
Multiple R-squared:  0.08771, Adjusted R-squared:  0.08047
F-statistic: 12.11 on 1 and 126 DF, p-value: 0.000688

```

אחוז גירושין:

```
Call:
lm(formula = Age.at.1st.birth ~ Divorce.rates, data = dataset,
    x = TRUE, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-7.0618 -3.5689 -0.3492  3.4328  9.5025

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.72649    1.72334   12.607  <2e-16 ***
Divorce.rates  0.07505    0.04026    1.864   0.0646 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.064 on 126 degrees of freedom
Multiple R-squared:  0.02684,    Adjusted R-squared:  0.01911
F-statistic: 3.475 on 1 and 126 DF,  p-value: 0.06464
```

משטר:

```
Call:
lm(formula = Age.at.1st.birth ~ factor(regime), data = dataset,
    x = TRUE, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-7.3051 -3.1301  0.1874  3.2949  8.5111

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.4051     0.4229   60.071  <2e-16 ***
factor(regime)2  -0.9162     1.3884   -0.660   0.5106
factor(regime)3   2.8449     2.8370    1.003   0.3180
factor(regime)4  -2.9051     2.0283   -1.432   0.1547
factor(regime)5  -1.3301     1.4650   -0.908   0.3658
factor(regime)6  -3.2851     1.8240   -1.801   0.0742 .
factor(regime)7  -3.1506     1.2688   -2.483   0.0144 *
factor(regime)8   7.1949     3.9898    1.803   0.0738 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.967 on 120 degrees of freedom
Multiple R-squared:  0.1168,    Adjusted R-squared:  0.06524
F-statistic: 2.266 on 7 and 120 DF,  p-value: 0.03347

~ |
```

דת:

```
Call:
lm(formula = Age.at.1st.birth ~ factor(religion), data = dataset,
    x = TRUE, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-6.3558 -3.8058 -0.1682  3.3692  9.8094

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.829     1.537   15.503  <2e-16 ***
factor(religion)2  1.427     1.602    0.891   0.375
factor(religion)3  1.195     2.806    0.426   0.671
factor(religion)4  5.005     2.806    1.783   0.077 .
factor(religion)5  3.771     4.347    0.868   0.387
factor(religion)6  -0.148     1.692   -0.087   0.930
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.067 on 122 degrees of freedom
Multiple R-squared:  0.05657,    Adjusted R-squared:  0.01791
F-statistic: 1.463 on 5 and 122 DF,  p-value: 0.2068
```


13.6 נספח ו'- תוצאות האלגוריתמים

רגרסיה לפנים:

```
Step: AIC=212.82
AgeAtFirstBirth ~ Avg.marriage.age..woman + Avg.num.of.kids +
  Days.at.home.for.labor + Wage + regimeFactor + Avg.marriage.age..woman:regimeFactor
```

	Df	Sum of Sq	RSS	AIC
<none>			593.98	212.82
+ Life.expectancy..women	1	5.6480	588.33	213.62
+ Rate.of.happiness	1	0.2165	593.77	214.77
+ Yrs.of.education..woman	1	0.1945	593.79	214.78
+ Days.at.home.for.labor:regimeFactor	2	8.9186	585.06	214.93
+ regimeFactor:Wage	2	1.7014	592.28	216.46

```
> summary(fwd.model)
```

```
Call:
lm(formula = AgeAtFirstBirth ~ Avg.marriage.age..woman + Avg.num.of.kids +
  Days.at.home.for.labor + Wage + regimeFactor + Avg.marriage.age..woman:regimeFactor,
  data = datasetNew)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-5.1276 -1.3078 -0.2719  0.7018  8.1388
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.297e+01  2.210e+00   5.870 4.23e-08 ***
Avg.marriage.age..woman  4.900e-01  7.573e-02   6.471 2.42e-09 ***
Avg.num.of.kids    -5.125e-01  2.728e-01  -1.878  0.0628 .
Days.at.home.for.labor  8.298e-03  4.304e-03   1.928  0.0563 .
Wage             3.758e-05  2.051e-05   1.832  0.0695 .
regimeFactor2    -4.560e+00  5.646e+00  -0.808  0.4209
regimeFactor3     5.766e+00  3.719e+00   1.550  0.1238
Avg.marriage.age..woman:regimeFactor2  2.206e-01  2.644e-01   0.835  0.4056
Avg.marriage.age..woman:regimeFactor3 -3.011e-01  1.566e-01  -1.923  0.0569 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.263 on 116 degrees of freedom
Multiple R-squared:  0.7106,    Adjusted R-squared:  0.6906
F-statistic: 35.6 on 8 and 116 DF, p-value: < 2.2e-16
```

רגרסיה לאחור:

```
Step: AIC=211.77
AgeAtFirstBirth ~ Days.at.home.for.labor + regimeFactor + Life.expectancy..women +
  Avg.marriage.age..woman + regimeFactor:Avg.marriage.age..woman
```

	Df	Sum of Sq	RSS	AIC
<none>			598.50	211.76
- Days.at.home.for.labor	1	19.892	618.40	213.85
- regimeFactor:Avg.marriage.age..woman	2	39.052	637.56	215.67
- Life.expectancy..women	1	36.631	635.14	217.19

```
> summary(bw.model)
```

```
Call:
lm(formula = AgeAtFirstBirth ~ Days.at.home.for.labor + regimeFactor +
  Life.expectancy..women + Avg.marriage.age..woman + regimeFactor:Avg.marriage.age..woman,
  data = datasetNew)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7854	-1.3319	-0.1523	0.8088	7.7483

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.467580	2.081829	1.666	0.09846 .
Days.at.home.for.labor	0.008311	0.004214	1.972	0.05097 .
regimeFactor2	-4.977042	5.471461	-0.910	0.36488
regimeFactor3	7.642933	3.495070	2.187	0.03075 *
Life.expectancy..women	0.108996	0.040731	2.676	0.00852 **
Avg.marriage.age..woman	0.533411	0.066084	8.072	6.88e-13 ***
regimeFactor2:Avg.marriage.age..woman	0.239760	0.257514	0.931	0.35374
regimeFactor3:Avg.marriage.age..woman	-0.380903	0.147478	-2.583	0.01103 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.262 on 117 degrees of freedom
Multiple R-squared: 0.7084, Adjusted R-squared: 0.6909
F-statistic: 40.6 on 7 and 117 DF, p-value: < 2.2e-16

רגרסיה בצעדים:

Step: AIC=211.77

AgeAtFirstBirth ~ Days.at.home.for.labor + regimeFactor + Life.expectancy..women +
Avg.marriage.age..woman + regimeFactor:Avg.marriage.age..woman

	Df	Sum of Sq	RSS	AIC
<none>			598.50	211.76
+ Wage	1	5.793	592.71	212.55
+ Days.at.home.for.labor:regimeFactor	2	12.634	585.87	213.10
+ Avg.num.of.kids	1	2.935	595.57	213.15
+ Yrs.of.education..woman	1	2.416	596.09	213.26
+ Rate.of.happiness	1	0.119	598.39	213.74
- Days.at.home.for.labor	1	19.892	618.40	213.85
- regimeFactor:Avg.marriage.age..woman	2	39.052	637.56	215.67
- Life.expectancy..women	1	36.631	635.14	217.19

> summary(stepwise)

Call:

```
lm(formula = AgeAtFirstBirth ~ Days.at.home.for.labor + regimeFactor +  
Life.expectancy..women + Avg.marriage.age..woman + regimeFactor:Avg.marriage.age..woman,  
data = datasetNew)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.7854	-1.3319	-0.1523	0.8088	7.7483

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.467580	2.081829	1.666	0.09846 .
Days.at.home.for.labor	0.008311	0.004214	1.972	0.05097 .
regimeFactor2	-4.977042	5.471461	-0.910	0.36488
regimeFactor3	7.642933	3.495070	2.187	0.03075 *
Life.expectancy..women	0.108996	0.040731	2.676	0.00852 **
Avg.marriage.age..woman	0.533411	0.066084	8.072	6.88e-13 ***
regimeFactor2:Avg.marriage.age..woman	0.239760	0.257514	0.931	0.35374
regimeFactor3:Avg.marriage.age..woman	-0.380903	0.147478	-2.583	0.01103 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.262 on 117 degrees of freedom

Multiple R-squared: 0.7084, Adjusted R-squared: 0.6909

F-statistic: 40.6 on 7 and 117 DF, p-value: < 2.2e-16

13.7 נספח ז'- בדיקת הנחות המודל:

```
#Linear Model
ModelAfterStep <- lm(AgeAtFirstBirth~Days.at.home.for.labor
                     +Life.expectancy..women+Avg.marriage.age..woman*regimeFactor
                     ,datasetAfterStep)
suma <- summary(ModelAfterStep)

#Fixed errors
datasetAfterStep$predicted <- fitted(ModelAfterStep)#predicted values
datasetAfterStep$residuals <- residuals(ModelAfterStep)
se <- sqrt(var(datasetAfterStep$residuals))
datasetAfterStep$residualsFix <- (residuals(ModelAfterStep)/se)#fixed errors
plot(datasetAfterStep$residualsFix,xlab = "predictedValues",ylab = "FixedErrors")
abline(h=0)

#QQ plot:
qqnorm(datasetAfterStep$residualsFix)
abline(a=0,b=1)

#Histogram plot:
hist(datasetAfterStep$residualsFix)

#-----KS-----
ks.test(x=datasetAfterStep$residualsFix,y="pnorm",alternative = "two.sided",exact=NULL)

#-----F Test for Equal Variances-----
A <- datasetAfterStep[,1]
A<-sort(A)
third <- round(length(A)/3)
twothird <- round(length(A)*2/3)
thirdData <- A[1:third]
twothirdData <- A[twothird:length(A)]
var.test(x= thirdData, y= twothirdData, ratio = 1, alternative = c("two.sided"), conf.level = 0.95)
```

13.8 נספח ח'- שימוש המודל

```
rowArab <- which(datasetTotal$Country.Name == "United Arab Emirates")
datasetTotal$Country.Name[rowArab]
print(datasetAfterStep[rowArab:rowArab,1:length(datasetAfterStep)-1] ) # dataOf United Arab Emirates
```

```

#checkPizor:
boxcox(ModelAfterStep, lambda = seq(-7,5,0.1))
lambda <- 0
ModelAfterCox <- lm(log(AgeAtFirstBirth)~Days.at.home.for.labor
                    +Life.expectancy..women+Avg.marriage.age..woman*regimeFactor
                    ,datasetAfterStep)
sama <- summary(ModelAfterCox)

#Fixed errors
datasetAfterStep$predicted <- fitted(ModelAfterCox)#predicted values
datasetAfterStep$predicted <- exp(datasetAfterStep$predicted)
datasetAfterStep$residuals <- datasetAfterStep$AgeAtFirstBirth-datasetAfterStep$predicted
se <- sqrt(var(datasetAfterStep$residuals))
datasetAfterStep$residualsFix <- (datasetAfterStep$residuals/se)#fixed errors
plot(datasetAfterStep$residualsFix,xlab = "predictedValues",ylab = "FixedErrors")
abline(h=0)

#findRajusted:
SSE <- sum((datasetAfterStep$residuals)^2)
RAadj <- 1-(SSE/(125-8))/var(datasetAfterStep$AgeAtFirstBirth)
RAadj

#QQ plot:
qqnorm(datasetAfterStep$residualsFix);axis(2,seq(1,4,1))
abline(a=0,b=1)

#Histogram plot:
hist(datasetAfterStep$residualsFix)

#-----KS
ks.test(x=datasetAfterStep$residualsFix,y="pnorm",alternative = "two.sided",exact=NULL)

#try another transforms(0.5):

ModelAfterCox <- lm(((AgeAtFirstBirth)^0.5)~Days.at.home.for.labor
                    +Life.expectancy..women+Avg.marriage.age..woman*regimeFactor
                    ,datasetAfterStep)
sama <- summary(ModelAfterCox)

#Fixed errors
datasetAfterStep$predicted <- fitted(ModelAfterCox)#predicted values
datasetAfterStep$predicted <- (datasetAfterStep$predicted)^(2)
datasetAfterStep$residuals <- datasetAfterStep$AgeAtFirstBirth-datasetAfterStep$predicted
se <- sqrt(var(datasetAfterStep$residuals))
datasetAfterStep$residualsFix <- (datasetAfterStep$residuals/se)#fixed errors
plot(datasetAfterStep$residualsFix,xlab = "predictedValues",ylab = "FixedErrors")
abline(h=0)

#findRajusted:
SSE <- sum((datasetAfterStep$residuals)^2)
RAadj <- 1-(SSE/(125-8))/var(datasetAfterStep$AgeAtFirstBirth)
RAadj

#QQ plot:
qqnorm(datasetAfterStep$residualsFix);axis(2,seq(1,4,1))
abline(a=0,b=1)

#Histogram plot:
hist(datasetAfterStep$residualsFix)

#Shapiro test
shapiro.test(datasetAfterStep$residualsFix)

#-----KS
ks.test(x=datasetAfterStep$residualsFix,y="pnorm",alternative = "two.sided",exact=NULL)

```

```

#try another transforms(-0.5):

ModelAfterCox <- lm(((AgeAtFirstBirth)^-0.5)~Days.at.home.for.labor
                    +Life.expectancy..women+Avg.marriage.age..woman*regimeFactor
                    ,datasetAfterStep)
sama <- summary(ModelAfterCox)

#Fixed errors
datasetAfterStep$predicted <- fitted(ModelAfterCox)#predicted values
datasetAfterStep$predicted <- (datasetAfterStep$predicted)^(-2)
datasetAfterStep$residuals <- datasetAfterStep$AgeAtFirstBirth-datasetAfterStep$predicted
se <- sqrt(var(datasetAfterStep$residuals))
datasetAfterStep$residualsFix <- (datasetAfterStep$residuals/se)#fixed errors
plot(datasetAfterStep$residualsFix,xlab = "predictedValues",ylab = "FixedErrors")
abline(h=0)

#findRajusted:
SSE <- sum((datasetAfterStep$residuals)^2)
RAdj <- 1-(SSE/(125-8))/var(datasetAfterStep$AgeAtFirstBirth)
RAdj

#QQ plot:
qqnorm(datasetAfterStep$residualsFix);axis(2,seq(1,4,1))
abline(a=0,b=1)

#Histogram plot:
hist(datasetAfterStep$residualsFix)

#-----KS
ks.test(x=datasetAfterStep$residualsFix,y="pnorm",alternative = "two.sided",exact=NULL)

```

13.9.1 נספח י'- קוד המודל

```
datasetTotal<-read.csv(file.choose(),header = T) #continuous variables without countries
colnames(datasetTotal)[1] <- "AgeAtFirstBirth"

dataset <- datasetTotal[,1:length(datasetTotal)-1] #remove country name

colnames(dataset)[1] <- "AgeAtFirstBirth"

dataWithoutCat <- subset(dataset,select=c("AgeAtFirstBirth", "Days.at.home.for.labor",
    "Rate.of.happiness",
    "Yrs.of.education..woman", "Divorce.rates", "Life.expectancy..women",
    "Hrs.of.work", "Avg.marriage.age..woman",
    "Avg.num.of.kids", "Wage"))

#-----Q3-----
cov(dataWithoutCat)
x <- cor(dataWithoutCat)

plot(dataWithoutCat,col=rgb(0,100,0,50,maxColorValue=255), pch=16)
summary(dataWithoutCat)

#-----Q4-----

library(e1071)

answer <- function(vector,string){ #to display data comfortably
  MEAN <- mean(vector)
  MEDIAN <- median(vector)
```

```

SD <- sd(vector)

QUANTILE <- quantile(vector,probs=c(0.25,0.5,0.75))

SKEWNESS <- skewness(vector)

print(string)

paste("mean=",MEAN," median=",MEDIAN," sd=",SD," quantile=", QUANTILE,"
skewness=",SKEWNESS)

}

print(answer(dataWithoutCat$AgeAtFirstBirth,"AgeAtFirstBirth"))
print(answer(dataWithoutCat$Days.at.home.for.labor,"Days.at.home.for.labor"))
print(answer(dataWithoutCat$Rate.of.happiness,"Rate.of.happiness"))
print(answer(dataWithoutCat$Yrs.of.education..woman,"Yrs.of.education..woman"))
print(answer(dataWithoutCat$Divorce.rates,"Divorce.rates") )
print(answer(dataWithoutCat$Life.expectancy..women,"expectancy..women"))
print(answer(dataWithoutCat$Hrs.of.work,"Hrs.of.work") )
print(answer(dataWithoutCat$Avg.marriage.age..woman,"Avg.marriage.age..woman"))
print(answer(dataWithoutCat$Avg.num.of.kids,"Avg.num.of.kids"))
print(answer(dataWithoutCat$Wage,"wage"))

categoricalDataset <- subset(dataset,select=c("AgeAtFirstBirth","regime"))

yAtRegime1 <- sqldf("select AgeAtFirstBirth from categoricalDataset where
regime=='Democracy'")

print(answer(yAtRegime1$AgeAtFirstBirth,"AgeAtFirstBirth"))

yAtRegime2 <- sqldf("select AgeAtFirstBirth from categoricalDataset where
regime='Dominant Party'")

print(answer(yAtRegime2$AgeAtFirstBirth,"AgeAtFirstBirth"))

yAtRegime3 <- sqldf("select AgeAtFirstBirth from categoricalDataset where
regime='Foreign/Occupied'")

```

```
print(answer(yAtRegime3$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtRegime4 <- sqldf("select AgeAtFirstBirth from categorialDataset where  
regime='Military'")
```

```
print(answer(yAtRegime4$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtRegime5 <- sqldf("select AgeAtFirstBirth from categorialDataset where  
regime='Monarchy'")
```

```
print(answer(yAtRegime5$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtRegime6 <- sqldf("select AgeAtFirstBirth from categorialDataset where regime='Party-  
Personal'")
```

```
print(answer(yAtRegime6$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtRegime7 <- sqldf("select AgeAtFirstBirth from categorialDataset where regime='Personal  
Dictatorship'")
```

```
print(answer(yAtRegime7$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtRegime8 <- sqldf("select AgeAtFirstBirth from categorialDataset where  
regime='Provisional - Civilian'")
```

```
print(answer(yAtRegime8$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
categorialDataset2<-subset(dataset,select=c("AgeAtFirstBirth","Religion"))
```

```
yAtReligion1 <- sqldf("select AgeAtFirstBirth from categorialDataset2 where  
Religion='Buddhist'")
```

```
print(answer(yAtReligion1$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtReligion2 <- sqldf("select AgeAtFirstBirth from categorialDataset2 where  
Religion='Christian'")
```

```
print(answer(yAtReligion2$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtReligion3 <- sqldf("select AgeAtFirstBirth from categorialDataset2 where Religion='Hindu'")
```

```
print(answer(yAtReligion3$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtReligion4 <- sqldf("select AgeAtFirstBirth from categorialDataset2 where Religion='IrReligion'")
```

```
print(answer(yAtReligion1$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtReligion5 <- sqldf("select AgeAtFirstBirth from categorialDataset2 where Religion='Jewish'")
```

```
print(answer(yAtReligion5$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
yAtReligion6 <- sqldf("select AgeAtFirstBirth from categorialDataset2 where Religion='Muslim'")
```

```
print(answer(yAtReligion6$AgeAtFirstBirth,"AgeAtFirstBirth"))
```

```
#-----Q5-----
```

```
bp<-boxplot(dataset$Days.at.home.for.labor , main='Days.at.home.for.labor')
```

```
bp<-boxplot(dataset$Rate.of.happiness , main='Rate.of.happiness')
```

```
bp<-boxplot(dataset$Yrs.of.education , main='Yrs.of.education')
```

```
bp<-boxplot(dataset$Divorce.rates , main='Divorce.rates')
```

```
bp<-boxplot(dataset$Life.expectancy , main='Life.expectancy')
```

```
bp<-boxplot(dataset$Hrs.of.work , main='Hrs.of.work')
```

```
bp<-boxplot(dataset$Avg.marriage.age , main='Avg.marriage.age')
```

```
bp<-boxplot(dataset$Avg.num.of.kids , main='Avg.num.of.kids')
```

```
bp<-boxplot(dataset$AgeAtFirstBirth , main='AgeAtFirstBirth')
```

```
bp<-boxplot(dataset$Wage , main='Wage')
```



```
#----- Q6-----
```

```
#-----daysAtHome-----
```

```
hist(dataset$Days.at.home.for.labor,prob=TRUE, main='Days at home for labor',xlab = 'Days at home for labor')
```

```
lines(density(dataset$Days.at.home.for.labor),col="blue",lwd=2)
```

```
plot(ecdf(dataset[, "Days.at.home.for.labor"]),main='Days at home for labor', xlab = 'Days at home for labor')
```

```
#-----AvNumOfKids-----
```

```
hist(dataset$Avg.num.of.kids,prob=TRUE, main='Avg hum of kids',xlab = 'Avg hum of kids')
```

```
lines(density(dataset$Avg.num.of.kids),col="blue",lwd=2)
```

```
plot(ecdf(dataset[, "Avg.num.of.kids"]),main='Avg.num.of.kids',xlab = 'Avg.num.of.kids')
```

```
#-----LifeExpectancy-----
```

```
hist(dataset$Life.expectancy,prob=TRUE, main='Life expectancy',xlab = 'Life expectancy')
```

```
lines(density(dataset$Life.expectancy),col="blue",lwd=2)
```

```
plot(ecdf(dataset[, "Life.expectancy..women"]),main='Life expectancy',xlab = 'Life expectancy')
```

```
# ----- Q7-----
```

```
plot(x=dataset$Rate.of.happiness,y=dataset$Life.expectancy,
```

```
      ylab="Life expectancy",frame = FALSE,col = 'blue' ,pch = 20,xlab="Rate of happiness")
```

```
plot(x=as.numeric(factor(dataset$Religion)),y=dataset$Wage,
```

```

ylab="Wage",frame = FALSE,col = 'blue',xlab="Religion")
plot(x=dataset$Yrs.of.education,y=dataset$Avg.num.of.kids ,
     ylab="Avg num of kids",frame = FALSE,col = 'blue' ,pch = 20,xlab="Years of education")
plot(x=dataset$Avg.marriage.age,y=dataset$Divorce.rates,
     ylab="Divorce rates",frame = FALSE,col = 'blue' ,pch = 20,xlab="Avg marriage age")
plot(x=dataset$Yrs.of.education,y=dataset$Wage,
     ylab="Wage",frame = FALSE,col = 'blue' ,pch = 20,xlab="Years of education")

```

```

#-----Q8-----
binFreqTable <- function(x, bins) {
  freq = hist(x, breaks=bins, include.lowest=TRUE, plot=FALSE)
  ranges = paste(head(freq$breaks,-1), freq$breaks[-1], sep=" - ")
  return(data.frame(range = ranges, frequency = freq$counts))
}

```

#1D tables

```

x2.9 <- binFreqTable(dataset$Rate.of.happiness,seq(2,8,by=1))
hist(dataset$Rate.of.happiness); axis(1,at=seq(2,8,1))

```

```

xx2.9 <- binFreqTable(dataset$Avg.num.of.kids,seq(0,7,by=1))
hist(dataset$Avg.num.of.kids); axis(1,at=seq(0,7,1))

```

#2D tables

#2D table- wage&hrs of work

```
ANS <- matrix( 0,nrow = 7, ncol = 7)
```

```
x=20
```

```
y=0
```

```
for(n in 1:7){
```

```
  for(k in 1:7){
```

```
    for(i in 1:128){
```

```
      if((dataWithoutCat[i,7]>=x && dataWithoutCat[i,7]<(5+x)) && (dataWithoutCat[i,10]>=y  
&& dataWithoutCat[i,10]<(10000+y))){
```

```
        ANS[n,k]=ANS[n,k]+1
```

```
      }
```

```
    }
```

```
    y=y+10000
```

```
  }
```

```
  y=0
```

```
  x=x+5
```

```
}
```

#2D table- avg age of marriage \$avg num of kids

```
ANS2 <- matrix( 0,nrow = 7, ncol = 7)
```

```
x=14
```

```
y=0
```

```
for(n in 1:7){
```

```
  for(k in 1:7){
```

```
    for(i in 1:128){
```

```
      if((dataWithoutCat[i,8]>=x && dataWithoutCat[i,8]<(3+x)) && (dataWithoutCat[i,9]>=y  
&& dataWithoutCat[i,9]<(1+y))){
```

```
        ANS2[n,k]=ANS2[n,k]+1
```

```
      }
```

```

    }
    y=y+1
  }
  y=0
  x=x+3
}

```

```

#-----Part B-----

```

```

#-----Q 2.1-----

```

```

#scatterplot for suspicious variables

```

```

plot(y=dataset$AgeAtFirstBirth,x=dataset$Days.at.home.for.labor,
      xlab="Days.at.home.for.labor",frame = TRUE,col = 'blue' ,pch = 20,ylab="AgeAtFirstBirth")
abline(lm(dataset$AgeAtFirstBirth~dataset$Days.at.home.for.labor), col="red")

```

```

plot(y=dataset$AgeAtFirstBirth,x=dataset$Divorce.rates,
      xlab="Divorce.rates",frame = TRUE,col = 'blue' ,pch = 20,ylab="AgeAtFirstBirth")
abline(lm(dataset$AgeAtFirstBirth~dataset$Divorce.rates), col="red")

```

```

plot(y=dataset$AgeAtFirstBirth,x=dataset$Hrs.of.work,
      xlab="Hrs.of.work",frame = TRUE,col = 'blue' ,pch = 20,ylab="AgeAtFirstBirth")
abline(lm(dataset$AgeAtFirstBirth~dataset$Hrs.of.work), col="red")

```

```

model.1<-lm(AgeAtFirstBirth ~ factor(Religion), data=dataset,x=TRUE,y=TRUE)
summary(model.1)

```

```
model.2<-lm(AgeAtFirstBirth ~ factor(regime), data=dataset,x=TRUE,y=TRUE)
```

```
summary(model.2)
```

```
model.3<-lm(AgeAtFirstBirth ~ Hrs.of.work, data=dataset,x=TRUE,y=TRUE)
```

```
summary(model.3)
```

```
model.4<-lm(AgeAtFirstBirth ~ Days.at.home.for.labor, data=dataset,x=TRUE,y=TRUE)
```

```
summary(model.4)
```

```
model.5<-lm(AgeAtFirstBirth ~ Divorce.rates, data=dataset,x=TRUE,y=TRUE)
```

```
summary(model.5)
```

```
datasetNew <- subset(dataset,select=c("AgeAtFirstBirth",      "Days.at.home.for.labor",  
                                     "Rate.of.happiness",  
                                     "Yrs.of.education..woman","regime",      "Life.expectancy..women",  
                                     "Avg.marriage.age..woman",      "Avg.num.of.kids",      "Wage"))
```

```
#-----Q2.2-----
```

```
datasetNew<-datasetNew[!(datasetNew$regime=="Provisional - Civilian"),]
```

```
datasetNew<-datasetNew[!(datasetNew$regime=="Foreign/Occupied"),]
```

```
unic <- unique(datasetNew$regime)
```

```
unic2 <- as.numeric(factor(unic))
```

```
datasetNew$regime <- as.numeric(factor(datasetNew$regime))
```

```
#1:democracy 2:dominant 3:millatery 4:monarchy 5:party-personal 6:dictator
```

```
#union categories
```

```
datasetNew$regime[datasetNew$regime==2] <- 1 #democracy and dominant
```

```
datasetNew$regime[datasetNew$regime==6] <- 2 #dictator
```

```
datasetNew$regime[datasetNew$regime==3] <- 3 #other
```

```
datasetNew$regime[datasetNew$regime==4] <- 3
```

```
datasetNew$regime[datasetNew$regime==5] <- 3
```

```
#discretion:
```

```
aveDays <- mean(datasetNew$Rate.of.happiness)
```

```
datasetNew$Rate.of.happiness<-ifelse(datasetNew$Rate.of.happiness>aveDays,c(0),c(1))
```

```
catRateHappiness<-lm(AgeAtFirstBirth ~ Rate.of.happiness,  
data=datasetNew,x=TRUE,y=TRUE)
```

```
summary(catRateHappiness)
```

```
#-----Q2.4-----
```

```
regimeFactor <- relevel(factor(datasetNew$regime),ref = c(1)) %>% print()
```

```
model.f<-lm(AgeAtFirstBirth ~ Avg.marriage.age..woman * regimeFactor, data =  
datasetNew)
```

```
plot(x=datasetNew$Avg.marriage.age..woman,y=datasetNew$AgeAtFirstBirth,col=regimeFactor)
```

```
lines(datasetNew$Avg.marriage.age..woman[regimeFactor==1],predict(model.f)[regimeFactor==1],col=1)
```

```
lines(datasetNew$Avg.marriage.age..woman[regimeFactor==2],predict(model.f)[regimeFactor==2],col=2)
```

```
lines(datasetNew$Avg.marriage.age..woman[regimeFactor==3],predict(model.f)[regimeFactor==3],col=3)
```

```
legend("topleft",legend=c("Democracy and  
Dominant","Dictatory","Else"),col=c(1,2,3),lty=c(1,1,1),bty="n",pt.bg=factor(regimeFactor))
```

```
summary(model.f)
```

```
model.f<-lm(AgeAtFirstBirth ~ Yrs.of.education..woman * regimeFactor, data = datasetNew)
```

```
plot(x=datasetNew$Yrs.of.education..woman,y=datasetNew$AgeAtFirstBirth,col=regimeFactor)
```

```
lines(datasetNew$Yrs.of.education..woman[regimeFactor==1],predict(model.f)[regimeFactor==1],col=1)
```

```
lines(datasetNew$Yrs.of.education..woman[regimeFactor==2],predict(model.f)[regimeFactor==2],col=2)
```

```
lines(datasetNew$Yrs.of.education..woman[regimeFactor==3],predict(model.f)[regimeFactor==3],col=3)
```

```
legend("topleft",legend=c("Democracy and  
Domimant","Dictatory","Else"),col=c(1,2,3),lty=c(1,1,1),bty="n",pt.bg=factor(regimeFactor))
```

```
summary(model.f)
```

```
model.f<-lm(AgeAtFirstBirth ~ Life.expectancy..women * regimeFactor, data = datasetNew)
```

```
plot(x=datasetNew$Life.expectancy..women,y=datasetNew$AgeAtFirstBirth,col=regimeFactor)
```

```
lines(datasetNew$Life.expectancy..women[regimeFactor==1],predict(model.f)[regimeFactor==1],col=1)
```

```
lines(datasetNew$Life.expectancy..women[regimeFactor==2],predict(model.f)[regimeFactor==2],col=2)
```

```
lines(datasetNew$Life.expectancy..women[regimeFactor==3],predict(model.f)[regimeFactor==3],col=3)
```

```
legend("topleft",legend=c("Democracy and  
Domimant","Dictatory","Else"),col=c(1,2,3),lty=c(1,1,1),bty="n",pt.bg=factor(regimeFactor))
```

```
summary(model.f)
```

```
model.f<-lm(AgeAtFirstBirth ~ Rate.of.happiness * regimeFactor, data = datasetNew)
```

```
plot(x=datasetNew$Rate.of.happiness,y=datasetNew$AgeAtFirstBirth,col=regimeFactor)
```

```
lines(datasetNew$Rate.of.happiness[regimeFactor==1],predict(model.f)[regimeFactor==1],col=1)
```

```
lines(datasetNew$Rate.of.happiness[regimeFactor==2],predict(model.f)[regimeFactor==2],col=2)
```

```
lines(datasetNew$Rate.of.happiness[regimeFactor==3],predict(model.f)[regimeFactor==3],col=3)
```

```
legend("topleft",legend=c("Democracy and  
Domimant","Dictatory","Else"),col=c(1,2,3),lty=c(1,1,1),bty="n",pt.bg=factor(regimeFactor))
```

```
summary(model.f)
```

```
model.f<-lm(AgeAtFirstBirth ~ Wage * regimeFactor, data = datasetNew)
```

```

plot(x=datasetNew$Wage,y=datasetNew$AgeAtFirstBirth,col=regimeFactor)

lines(datasetNew$Wage[regimeFactor==1],predict(model.f)[regimeFactor==1],col=1)

lines(datasetNew$Wage[regimeFactor==2],predict(model.f)[regimeFactor==2],col=2)

lines(datasetNew$Wage[regimeFactor==3],predict(model.f)[regimeFactor==3],col=3)

legend("topleft",legend=c("Democracy and
Domimant","Dictatory","Else"),col=c(1,2,3),lty=c(1,1,1),bty="n",pt.bg=factor(regimeFactor))

summary(model.f)

```

```

model.f<-lm(AgeAtFirstBirth ~ Days.at.home.for.labor * regimeFactor, data = datasetNew)

plot(x=datasetNew$Days.at.home.for.labor,y=datasetNew$AgeAtFirstBirth,col=regimeFactor)

lines(datasetNew$Days.at.home.for.labor[regimeFactor==1],predict(model.f)[regimeFactor==1],col=1)

lines(datasetNew$Days.at.home.for.labor[regimeFactor==2],predict(model.f)[regimeFactor==2],col=2)

lines(datasetNew$Days.at.home.for.labor[regimeFactor==3],predict(model.f)[regimeFactor==3],col=3)

legend("topleft",legend=c("Democracy and
Domimant","Dictatory","Else"),col=c(1,2,3),lty=c(1,1,1),bty="n",pt.bg=factor(regimeFactor))

summary(model.f)

```

```

model.f<-lm(AgeAtFirstBirth ~ Avg.num.of.kids * regimeFactor, data = datasetNew)

plot(x=datasetNew$Avg.num.of.kids,y=datasetNew$AgeAtFirstBirth,col=regimeFactor)

lines(datasetNew$Avg.num.of.kids[regimeFactor==1],predict(model.f)[regimeFactor==1],col=1)

lines(datasetNew$Avg.num.of.kids[regimeFactor==2],predict(model.f)[regimeFactor==2],col=2)

lines(datasetNew$Avg.num.of.kids[regimeFactor==3],predict(model.f)[regimeFactor==3],col=3)

legend("topleft",legend=c("Democracy and
Domimant","Dictatory","Else"),col=c(1,2,3),lty=c(1,1,1),bty="n",pt.bg=factor(regimeFactor))

summary(model.f)

```

#-----Q3.1-----


```

regimeFactor

fullModel <-
lm(AgeAtFirstBirth~Days.at.home.for.labor+Days.at.home.for.labor*regimeFactor+Rate.of.h
appiness+Yrs.of.education..woman

+Life.expectancy..women+Avg.marriage.age..woman+Avg.marriage.age..woman*regimeFact
or+Avg.num.of.kids

+Wage+Wage*regimeFactor,datasetNew)

```

```

summary(fullModel)

```

```

emptyModel <- lm(AgeAtFirstBirth~1,datasetNew)

```

```

#regration by steps:

```

```

bw.model <- step(fullModel,direction='backward',scope=~1)

```

```

summary(bw.model)

```

```

fwd.model <- step(emptyModel,direction='forward',scope=formula(fullModel))

```

```

summary(fwd.model)

```

```

stepwise <- step(fullModel, direction="both")

```

```

summary(stepwise)

```

```

#-----Q3.2-----

```

```

#delete Years Of Education:

```

```

datasetAfterStep <- datasetNew

```

```

datasetAfterStep$Rate.of.happiness <- NULL

```

```

datasetAfterStep$Yrs.of.education..woman <- NULL

```

```

datasetAfterStep$Avg.num.of.kids <- NULL

```

```

datasetAfterStep$Wage <- NULL

```

```

#Linear Model

```

```

ModelAfterStep <- lm(AgeAtFirstBirth~Days.at.home.for.labor
                     +Life.expectancy..women+Avg.marriage.age..woman*regimeFactor
                     ,datasetAfterStep)

suma <- summary(ModelAfterStep)

#Fixed errors
datasetAfterStep$predicted <- fitted(ModelAfterStep)#predicted values
datasetAfterStep$residuals <- residuals(ModelAfterStep)
se <- sqrt(var(datasetAfterStep$residuals))
datasetAfterStep$residualsFix <- (residuals(ModelAfterStep)/se)#fixed errors
plot(datasetAfterStep$residualsFix,xlab = "predictedValues",ylab = "FixedErrors")
abline(h=0)

#QQ plot:

qqnorm(datasetAfterStep$residualsFix)
abline(a=0,b=1)

#Histogram plot:
hist(datasetAfterStep$residualsFix)

#-----KS-----
ks.test(x=datasetAfterStep$residualsFix,y="pnorm",alternative = "two.sided",exact=NULL)

#-----F Test for Equal Variances-----

A <- datasetAfterStep[,1]

A<-sort(A)

```

```
third <- round(length(A)/3)
```

```
twothird <- round(length(A)*2/3)
```

```
thirdData <- A[1:third]
```

```
twothirdData <- A[twothird:length(A)]
```

```
var.test(x= thirdData, y= twothirdData, ratio = 1, alternative = c("two.sided"), conf.level =  
0.95)
```

```
#-----Q3.3-----
```

```
rowArab <- which(datasetTotal$Country.Name == "United Arab Emirates")
```

```
datasetTotal$Country.Name[rowArab]
```

```
print(datasetAfterStep[rowArab:rowArab,1:length(datasetAfterStep)-1] ) # dataOf United  
Arab Emirates
```

```
#-----Q4-----
```

```
#checkPizor:
```

```
boxcox(ModelAfterStep, lambda = seq(-7,5,0.1))
```

```
lambda <- 0
```

```
ModelAfterCox <- lm(log(AgeAtFirstBirth)~Days.at.home.for.labor
```

```
+Life.expectancy..women+Avg.marriage.age..woman*regimeFactor  
,datasetAfterStep)
```

```
sama <- summary(ModelAfterCox)
```

```
#Fixed errors
```

```
datasetAfterStep$predicted <- fitted(ModelAfterCox)#predicted values
```

```
datasetAfterStep$predicted <- exp(datasetAfterStep$predicted)
```

```
datasetAfterStep$residuals <- datasetAfterStep$AgeAtFirstBirth-datasetAfterStep$predicted
```

```
se <- sqrt(var(datasetAfterStep$residuals))
```

```
datasetAfterStep$residualsFix <- (datasetAfterStep$residuals/se)#fixed errors
```

```
plot(datasetAfterStep$residualsFix,xlab = "predictedValues",ylab = "FixedErrors")
```

```
abline(h=0)
```

```
#findRajusted:
```

```
SSE <- sum((datasetAfterStep$residuals)^2)
```

```
RAadj <- 1-(SSE/(125-8))/var(datasetAfterStep$AgeAtFirstBirth)
```

```
RAadj
```

```
#QQ plot:
```

```
qqnorm(datasetAfterStep$residualsFix);axis(2,seq(1,4,1))
```

```
abline(a=0,b=1)
```

```
#Histogram plot:
```

```
hist(datasetAfterStep$residualsFix)
```

```
#-----KS
```

```
ks.test(x=datasetAfterStep$residualsFix,y="pnorm",alternative = "two.sided",exact=NULL)
```

```
#-----
```

```
#try another transforms(0.5):
```

```
ModelAfterCox <- lm(((AgeAtFirstBirth)^0.5)~Days.at.home.for.labor  
+Life.expectancy..women+Avg.marriage.age..woman*regimeFactor  
,datasetAfterStep)
```

```
sama <- summary(ModelAfterCox)
```

```
#Fixed errors
```

```
datasetAfterStep$predicted <- fitted(ModelAfterCox)#predicted values  
datasetAfterStep$predicted <- (datasetAfterStep$predicted)^(2)  
datasetAfterStep$residuals <- datasetAfterStep$AgeAtFirstBirth-datasetAfterStep$predicted  
se <- sqrt(var(datasetAfterStep$residuals))  
datasetAfterStep$residualsFix <- (datasetAfterStep$residuals/se)#fixed errors  
plot(datasetAfterStep$residualsFix,xlab = "predictedValues",ylab = "FixedErrors")  
abline(h=0)
```

```
#findRajusted:
```

```
SSE <- sum((datasetAfterStep$residuals)^2)  
RAdj <- 1-(SSE/(125-8))/var(datasetAfterStep$AgeAtFirstBirth)  
RAdj
```

```
#QQ plot:
```

```
qqnorm(datasetAfterStep$residualsFix);axis(2,seq(1,4,1))
```

```
abline(a=0,b=1)
```

```
#Histogram plot:
```

```
hist(datasetAfterStep$residualsFix)
```

```
#Shapiro test
```

```
shapiro.test(datasetAfterStep$residualsFix)
```

```
#-----KS
```

```
ks.test(x=datasetAfterStep$residualsFix,y="pnorm",alternative = "two.sided",exact=NULL)
```

```
#-----
```

```
#try another transforms(-0.5):
```

```
ModelAfterCox <- lm(((AgeAtFirstBirth)^-0.5)~Days.at.home.for.labor  
+Life.expectancy..women+Avg.marriage.age..woman*regimeFactor  
,datasetAfterStep)  
sama <- summary(ModelAfterCox)
```

```
#Fixed errors
```

```
datasetAfterStep$predicted <- fitted(ModelAfterCox)#predicted values
```

```
datasetAfterStep$predicted <- (datasetAfterStep$predicted)^(-2)
```

```
datasetAfterStep$residuals <- datasetAfterStep$AgeAtFirstBirth-datasetAfterStep$predicted
```

```
se <- sqrt(var(datasetAfterStep$residuals))
datasetAfterStep$residualsFix <- (datasetAfterStep$residuals/se)#fixed errors
plot(datasetAfterStep$residualsFix,xlab = "predictedValues",ylab = "FixedErrors")
abline(h=0)
```

#findRajusted:

```
SSE <- sum((datasetAfterStep$residuals)^2)
RAdj <- 1-(SSE/(125-8))/var(datasetAfterStep$AgeAtFirstBirth)
RAdj
```

#QQ plot:

```
qqnorm(datasetAfterStep$residualsFix);axis(2,seq(1,4,1))
abline(a=0,b=1)
```

#Histogram plot:

```
hist(datasetAfterStep$residualsFix)
```

#-----KS

```
ks.test(x=datasetAfterStep$residualsFix,y="pnorm",alternative = "two.sided",exact=NULL)
```

#F Test for Equal Variances

```
A <- datasetAfterStep[,1]
```

```
A<-sort(A)
```

```
third <- round(length(A)/3)
```

```
twothird <- round(length(A)*2/3)
```

```
thirdData <- A[1:third]
```

```
twothirdData <- A[twothird:length(A)]
```

```
var.test(x= thirdData, y= twothirdData, ratio = 1, alternative = c("two.sided"), conf.level =  
0.95)
```