**פרויקט רגרסיה**

**חלק ב**



Cancer mortality rate

קבוצה 18

**315407114 | 208080879 | 209060334**

**תוכן עניינים:**

תקציר מנהלים .......................................................................................................................................3-4

הסרת משתנים ......................................................................................................................................5-6

התאמת משתנים ...................................................................................................................................7-8

משתני דמה...............................................................................................................................................9

משתני אינטראקציה .............................................................................................................................9-12

בחירת משתני המודל החדש .....................................................................................................................13

בדיקת הנחות בצורה ויזואלית ....................................................................................................................14

בדיקת הנחות בצורה פורמלית ..................................................................................................................15

שיפור המודל ......................................................................................................................................... 16

נספחים ............................................................................................................................................17-20

**1. תקציר מנהלים:**בפרויקט זה התמקדנו בניתוח ובחינה של המשתנים המשפיעים על שיעור התמותה מסרטן, במטרה לבנות מודל רגרסיה המסביר את הקשר בין המשתנים המסבירים לשיעור התמותה מסרטן. תהליך העבודה כלל איסוף נתונים רלוונטיים על המשתנים המסבירים, אשר משפיעים כל אחד באופן שונה על המשתנה המוסבר- אחוז התמותה מסרטן.

תחילה, בדקנו את השפעתם של המשתנים המסבירים על המוסבר באמצעות מקדם פירסון, רמת מובהקות וגרף הפיזור של כל אחד ביחס לשיעור התמותה מסרטן, בהתאם לבדיקות אלו הוחלט להסיר מן המודל את המשתנים Percentage of residents in citiesו-Depression rateאשר ניתן היה לראות כי מקדם פירסון היה נמוך, בעל מובהקות נמוכה וגרף הפיזור לא הראה קשר מסוים. בהמשך התהליך, בדקנו את הצורך בהמרת משתנים רציפים לקטגוריאליים ואיחוד קטגוריות במשתנים הקטגוריאליים. החלטנו לאחד קטגוריות תחת Air pollution index לשלוש קטגוריות במקום 5.פעולה זו סייעה לשיפור ההתאמה בין המודל לנתונים, ובכך הגדילה את דיוק המודל.לאחר מכן, יצרנו משתני דמה ומשתני אינטראקציה עבור המשתנה הקטגוריאלי יבשת לפי השערותינו לקשרים מעניינים. לבסוף, יצרנו את המודל לאחר כל השינויים ועברנו לשלב בו רצינו לבדוק מהו המודל הטוב ביותר. השתמשנו באלגוריתמים Forward selection ו- Backward selection על מנת לבחור את המודל הטוב ביותר, והשוונו ביניהם לפי מדד AICובדקנו גם את מדד BIC. לאחר פעולה זו קיבלנו מודל משופר, עבורו בדקנו האם הנחות המודל מתקיימות. בדקנו את הנחת הלינאריות, הנחת שוויון שונויות והנחת הנורמאליות. ביצענו בדיקה ויזואלית ופורמלית לבחינת קיום ההנחות. על פי הבדיקה הפורמלית הסקנו כי הנחת הלינאריות אינה מתקיימת וכי הנחת שוויון השונויות והנורמליות מתקיימות.   
לאחר מכן ביצענו טרנספורמציה על המודל על מנת שהוא יקיים את כל ההנחות. ביצענו טרנספורמציית לוג על המשתנה המוסבר Cancer Mortality Rate. כמו כן ביצענו טרנספורמציית שורש על Obesity Rate. לאחר ביצוע טרנספורמציות קיבלנו מודל משופר הן מבחינת מדדי הביצוע והן מבחינת עמידה בהנחות הרגרסיה. המודל המשופר הסופי הינו:

טבלת משתנים במודל:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **שם המשתנה** | **הסבר קצר על  המשתנה** | **סוג משתנה מוסבר/מסביר** | **סימון ב-R** | **יחידת מידה** | **סוג המשתנה -רציף/ קטגוריאלי** |
| **Country Index** | - | - | Country Index | - | - |
| **Cancer mortality rate** | שיעור התמותה ממחלת הסרטן | מוסבר | Cancer mortality rate | אחוז | רציף |
| **Obesity rate** | שיעור השמנת יתר | מסביר | Obesity rate | אחוז | רציף |
| **Smoke rate** | שיעור המעשנים | מסביר | Smoke rate | אחוז | רציף |
| **Air pollution index** | מדד זיהום אוויר | מסביר | Air pollution index | חסר יחידות | קטגוריאלי |
| **State Development Index** | מדד התפתחות במדינה | מסביר | State Development Index (0-1) | ללא יחידות | רציף |
| **Median age** | חציון הגיל | מסביר | Median age | שנים | רציף |
| **Average temperature** | טמפרטורה ממוצעת | מסביר | Average temperature | מעלות צלזיוס | רציף |
| **Percentage of residents in cities** | שיעור התושבים בעיר | מסביר | Percentage of residents in cities | אחוז | רציף |
| **Depression rate** | שיעור הלוקים בדיכאון | מסביר | Depression rate (%) | אחוז | רציף |

**2. עיבוד מקדים:**

* 1. **הסרה של משתנים:**

נרצה לבחון אילו משתנים מסבירים אינם משפיעים על אחוז התמותה מסרטן (המשתנה המוסבר). נבדוק זאת באמצעות מקדם פירסון המודד את הקשר בין שני משתנים רציפים ומתאר את חוזק הקשר ביניהם. ערכי המדד נעים בין (-1) ל1, ככל שערך המדד קרוב יותר לקצוות קיים קשר חזק יותר בין המשתנה המסביר למשתנה המוסבר, כשאר הקרבה ל1 מראה על קשר חיובי בין המשתנים וקרבה ל(-1) מראה על קשר שלילי בין המשתנים. כיוון שפירסון תקף רק על משתנים רציפים, לא נוכל לבדוק את השפעת המשתנים הקטגוריאליים: רמת זיהום האוויר ויבשות.

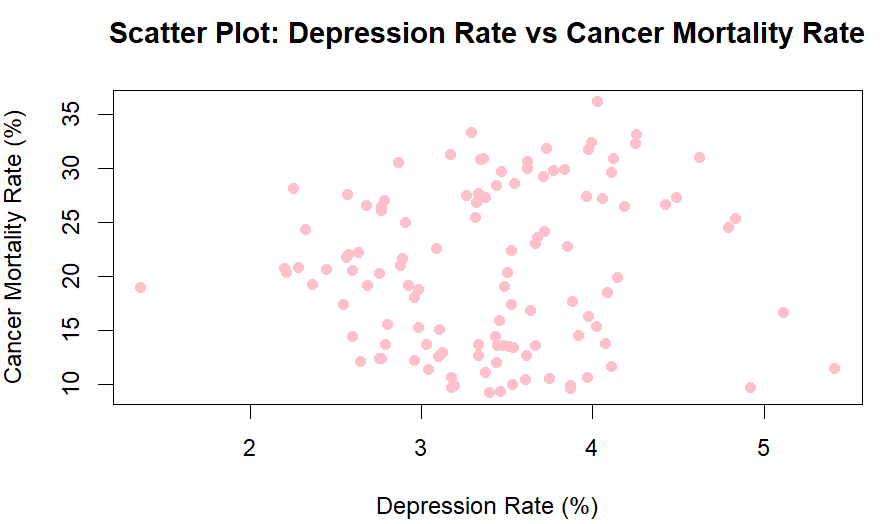
תוצאות הבדיקה הינן:

|  |  |
| --- | --- |
| משתנה מסביר | מקדם פירסון |
| Obesity rate | 0.23002 |
| Smoke rate | 0.25079 |
| State Development | 0.81869 |
| Median age | 0.73459 |
| Average temperature | 0.50409- |
| Percentage of residents in cities | 02034.0- |
| Depression rate | 08436.0 |

על פי מקדם פירסון, שני המשתנים האחרונים בטבלה מראים קשר חלש מאוד עם המשתנה המוסבר (מקדם קרוב לאפס), מה שמעיד על כך שתרומתם להסבר השונות של המשתנה המוסבר זניחה. על כן ארצה לבחון האם כדאי להסיר את 2 המשתנים הללו

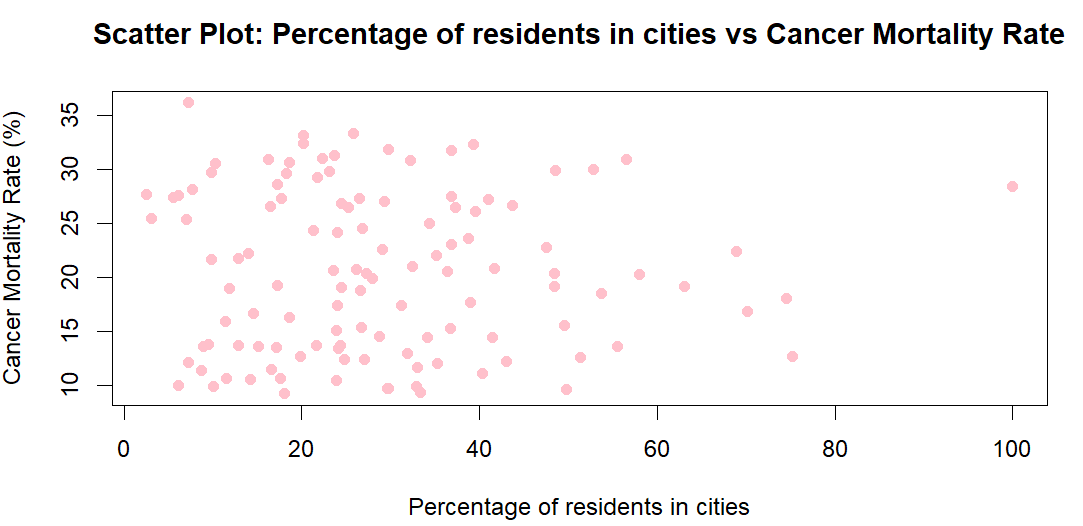
ארצה לבחון את תרשימי הפיזור של 2 המשתנים הללו על מנת לבחון האם כדאי להוריד אותם המודל:

תרשים פיזור עבור Depression rate:



הגרף מציג פיזור נקודות אקראי, ללא מגמה ברורה או קשר לינארי בין Depression Rateלבין Cancer Mortality Rate הפיזור האקראי מעיד על כך שהמשתנה Depression Rate אינו מסביר את השונות במשתנה המוסבר בצורה מובהקת. מידע זה מתסווף לכך שמקדם המתאם קרוב ל0. כתוצאה מכך, משתנה זה אינו תורם משמעותית למודל וניתן לשקול להסירו כדי לפשט את המודל.

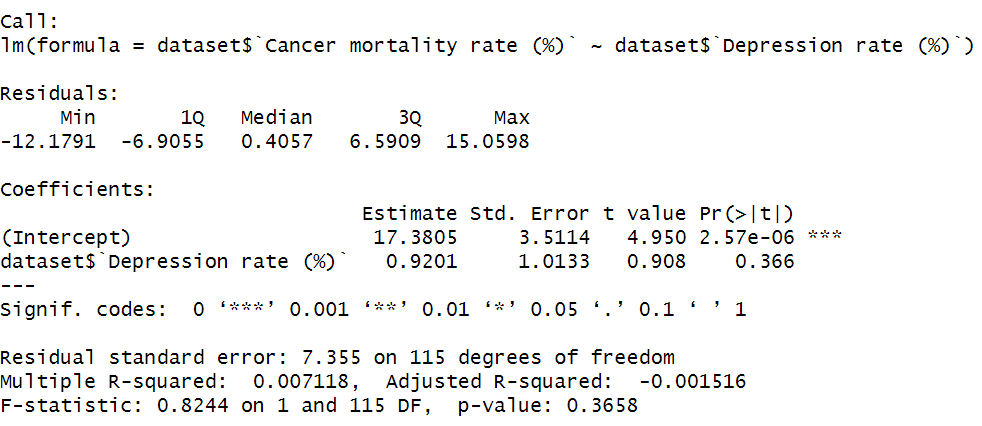
תרשים פיזור עבור Percentage of residents in cities:



מקדם פירסון של המשתנה קרוב ל0, דבר המעיד על קשר זניח מאוד עם המשתנה המוסבר. תרשים הפיזור מציג נקודות מפוזרות באופן אקראי ללא מגמה ברורה, מה שמחזק את הטענה שאין קשר לינארי או משמעותי בין המשתנים. בהתאם לכך, המשתנה אינו תורם להסבר השונות במודל וניתן להסירו כדי לפשט את המודל ולהפחית רעש סטטיסטי.

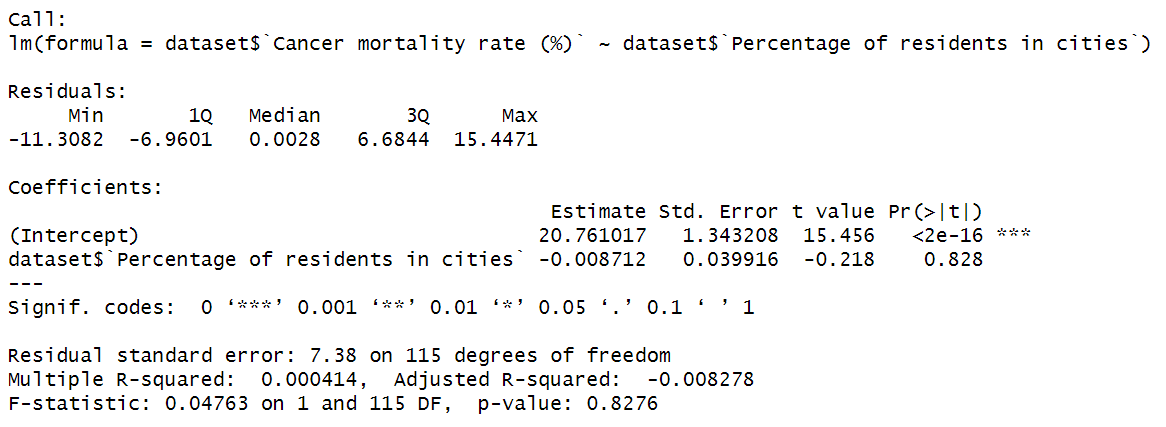
על מנת לוודא בצורה סופית את השפעת המשתנים על המודל, אבדוק את רמת המובהקות שלהם (P-val):

עבור המשתנה Depression rate:



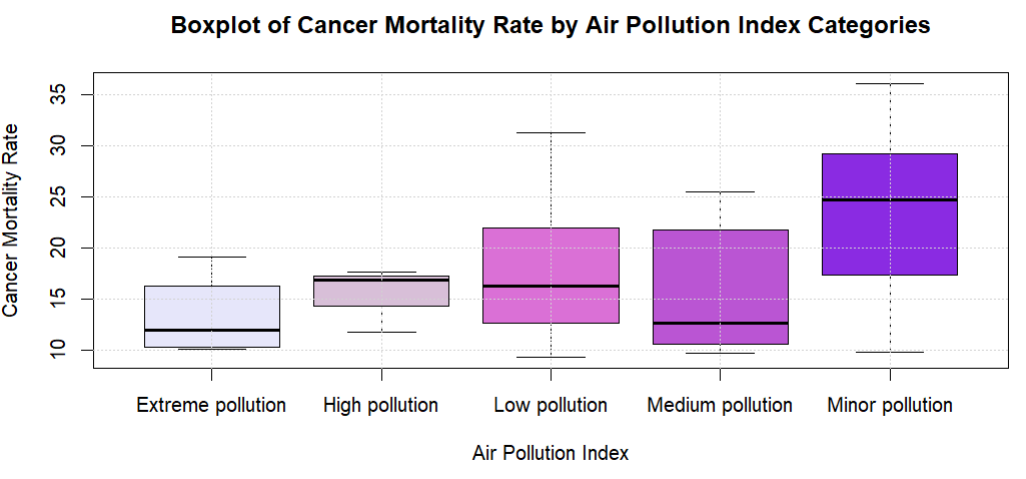
המשתנה Depression rate אינו מובהק סטטיסטית כמסביר עבור המשתנה המוסבר Cancer mortality rate וזאת בהתבסס על ה-p-value שהתקבל במודל, שערכו 0.366. משתנה נחשב למובהק סטטיסטית אם ה-p-value שלו קטן מ0.05. מאחר שהערך שהתקבל גדול משמעותית מ-0.05, אנו מסיקים כי הקשר בין המשתנים עשוי להיות מקרי ואינו משמעותי סטטיסטית.  
נוסיף זאת לממצאים שהצגנו קודם לכן, הקורלציה שמעידה על קשר חלש בין המשתנים וגרף הפיזור שלא הצביע על מגמה ברורה או קו מנחה כלשהו בקשר בין המשנים. בהתבסס על כל הסיבות הללו , נמליץ להוריד את המשתנה Depression rate מהמודל.

עבור Percentage of residents in cities:



המשתנה Percentage of residents in cities אינו מובהק סטטיסטית כמסביר עבור המשתנה המוסבר Cancer mortality rate וזאת על סמך ה-p-value שהתקבל במודל, שערכו 0.828. על פי כלל אצבע, משתנה נחשב למובהק סטטיסטית אם ה-p-value שלו קטן מ .0.05 מאחר שהערך שהתקבל גדול בהרבה מ-0.05, ניתן להסיק כי אין עדות סטטיסטית לכך שיש קשר מובהק בין המשתנים. בנוסף, ערך המתאם (correlation) בין המשתנים עמד על -0.02034 דבר המעיד על קשר כמעט אפסי וחלש מאוד בין המשתנים. גרף הפיזור שנוצר אף הוא לא הראה מגמה ברורה או קשר חזותי משמעותי בין המשתנים. בהתבסס על כל הסיבות הללו נמליץ להוריד את המשתנה Percentage of residents in citiesמהמודל.

**2.2 התאמת משתנים:**  
**2.2.1 בדיקת האופציה: איחוד קטגוריות במשתנה "Air Pollution Index" הקטגוריאלי**למשתנה "Air Pollution Index" חמש קטגוריות. נבדוק את הפיזור על ידי גרף Boxplot ונתונים סטטיסטיים:





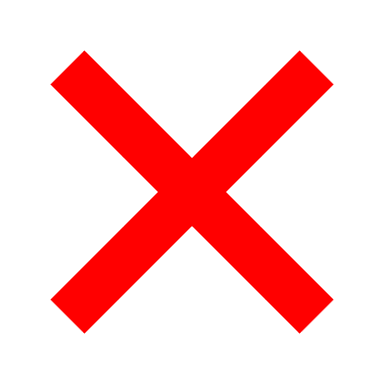
תמונה שמכילה טקסט, גופן, צילום מסך, מספר

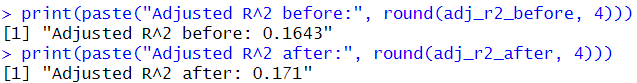
התיאור נוצר באופן אוטומטי



נבחן האם יש לבצע איחוד בין Extreme pollution ו-High pollution: מבחינת הפיזור, לשתי הקטגוריות יש פיזור מצומצם (Whiskers קצרים וקופסאות קטנות), שמראה שונות נמוכה בין הנתונים. הטווח של Extreme pollution נע בין 10 ל-19, ושל High pollution בין 11.7 ל-17.6 – יש חפיפה בין הטווחים. בנוסף, מספר תצפיות קטן (4 ו-3) מקשה על הבחנה משמעותית בין הקטגוריות. יש דמיון בממוצעים, וסטיות התקן בשתי הקטגוריות נמוכות, אך מספר התצפיות הקטן הופך את הנתונים לפחות מייצגים. לכן, האיחוד לקטגוריה אחת (High pollution) מוצדק.

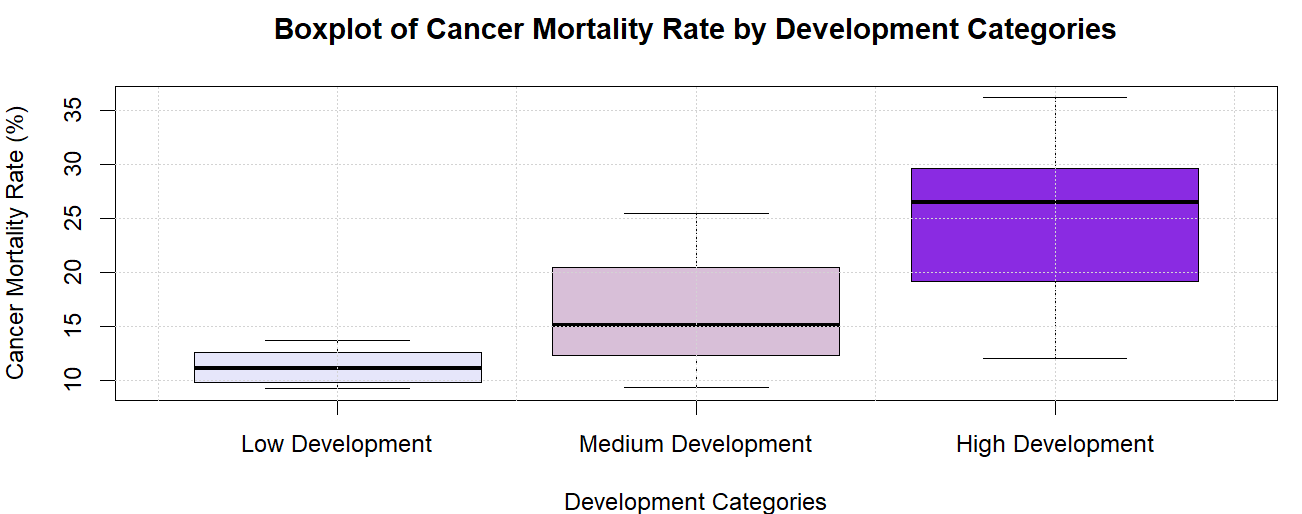
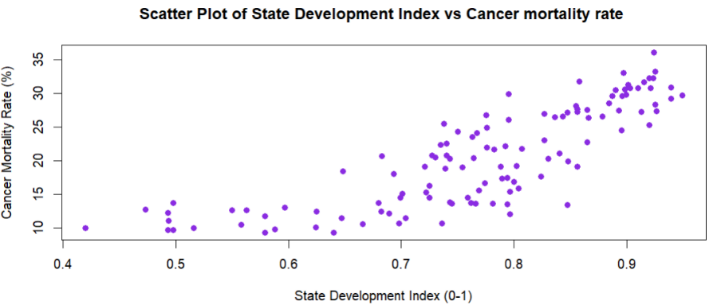
נבחן האם יש לבצע איחוד בין Low pollution ו-Medium pollution: הטווחים (Whiskers) של שתי הקטגוריות חופפים מאוד, במיוחד באזור הערכים הנמוכים (9–25). החציון (הקו שבתוך הקופסה) קרוב מאוד בין שתי הקטגוריות (16.33 לעומת 12.68).הקופסאות (IQR - Interquartile Range) מצביעות על דמיון בתפוצת הנתונים המרכזיים. לכן, נגיד שאין הבדל משמעותי בפיזור או במרכז הנתונים בין Low pollution ל-Medium pollution. לכן, איחוד הקטגוריות לקטגוריה אחת (Medium pollution) מוצדק.

נבחן האם כדאי לאחד את Minor pollution עם קטגוריה נוספת: הפיזור של Minor pollution (9.7–36.1) רחב משמעותית בהשוואה לקטגוריות אחרות. החציון (24.7) גבוה משמעותית משאר הקטגוריות. הקופסה (IQR) רחבה מאוד, מציינת שונות גבוהה בין מדינות בתוך הקטגוריה. סטיית התקן הגבוהה (7.03) גם מצביעה על שונות רחבה בתוך הקטגוריה. קטגוריה זו מכילה כמעט 60% מהתצפיות (כ-70 תצפיות), ולכן גם בגלל טיעון זה לא נרצה לאחד אותה. לכן, יש להשאיר קטגוריה זו בנפרד.

לסיכום, נשים לב כי מדד ה-Radj השתפר. נחליט שנאחד ונשאר עם 3 קטגוריות: High pollution, Medium pollution, Minor pollution

תמונה שמכילה טקסט, צילום מסך, תרשים, קו

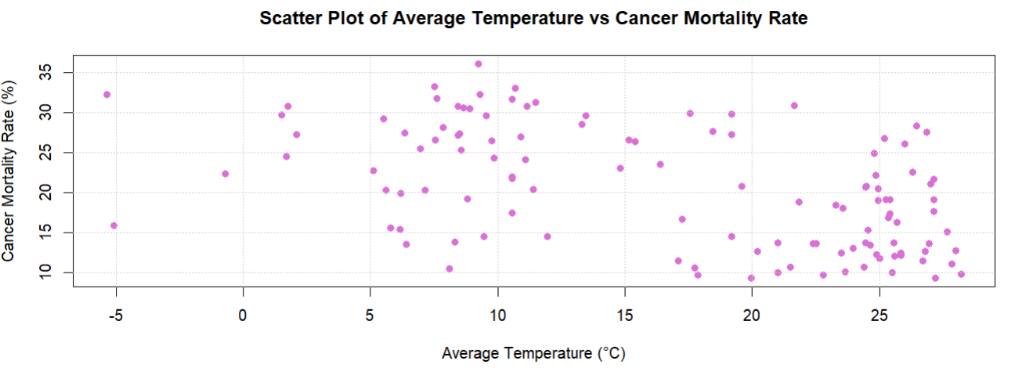
התיאור נוצר באופן אוטומטי

****תמונה שמכילה טקסט, גופן, צילום מסך, כחול חשמלי

התיאור נוצר באופן אוטומטי**2.2.2 בדיקת האופציה: המרת משתנה רציף "State Development Index" לקטגוריאלי:**על סמך הגרף, ניתן לראות מגמה חיובית בין מדד הפיתוח לבין שיעור התמותה מסרטן. חלוקת המשתנה לקטגוריות יכולה לפשט את הניתוח ולהבליט תבניות ברורות. עם זאת, כאשר המשתנה הוא רציף וגרף הפיזור מתאר התנהגות יחסית לינארית, אנו מנצלים את כל המידע שבו כרציף וכך נוכל לאמוד את הקשר בצורה מדויקת יותר .למרות ההבנה שככל הנראה לא יהיה שיפור בביצועי המודל ניסינו לבצע חלוקה הגיונית לקטגוריות. לאחר מכן, בדקנו את מדד הביצוע R^2 adj וראינו באופן ויזואלי שהוא לא השתפר.

לכן, נחליט **לא** לבצע את השינוי מתוך ההבנה שחלוקה לקטגוריות לא רלוונטית למשתנה זה ולא משפרת את ביצועי המודל.

**2.2.3 בדיקת האופציה: המרת משתנה רציף "Average Temperature" לקטגוריאלי**מבחינת פיזור, מהגרף ניתן לראות כי הטווח של הטמפרטורה הממוצעת נע בין כ-5 ל25 מעלות צלזיוס וכן שיעור התמותה מסרטן נע בין כ-10% ל-35%. הנתונים מפוזרים בצורה יחסית אחידה בטווח הטמפרטורות, ללא ריכוז ברור בטווח ספציפי.



על פי הגרף ישנה שונות גבוהה בנתונים בין שיעורי התמותה לנתוני ממוצע הטמפרטורות. לכן, לא ניתן לשייך את הנתונים לרמות מוגדרות מראש על בסיס קטגוריות, ונראה שאין הצדקה להפוך את המשתנה מטמפרטורה ממוצעת לרצף קטגוריאלי, להפך, האיחוד היה גורם לנו לאיבוד מידע חשוב וחוסר דיוק בנתונים, ולכן לא נבצע שום חלוקה לקטגוריות **מתוך הבנה שביצועי המודל לא ישתפרו ואין חלוקה הגיונית לקטגוריות.** לסיכום, נחליט ש**לא** להפוך את המשתנה הרציף Average Temperature לקטגוריאלי.

**2.3 הגדרת משתני דמה:**

המשתנה הקטגוריאלי שבחרנו הוא Continent עבורו נגדיר את משתני הדמה. משתנה זה הוא בעל *6 קטגוריות: Asia, Europe ,Africa ,Australia ,South America ,North America.*

*נגדיר את להיות יבשת כאשר - אירופה, - אפריקה, - אוסטרליה, - דרום אמריקה, - צפון אמריקה. בנוסף, נגדיר את קבוצת הבסיס להיות אסיה ובהתאם אליה נבנה את משתני הדמה:*

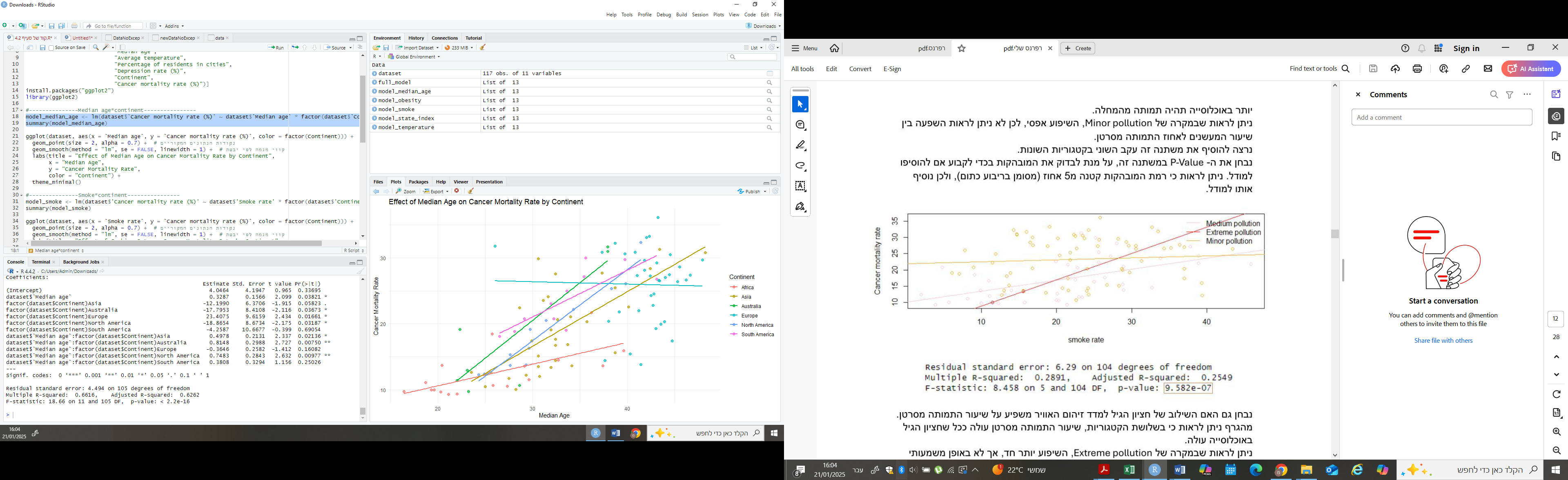
לצורכי הגדרת המודל המלא בסעיף 2.4, נגדיר גם את שאר משתני הדמה הנגזרים מהמשתנים הקטגוריאליים האחרים:

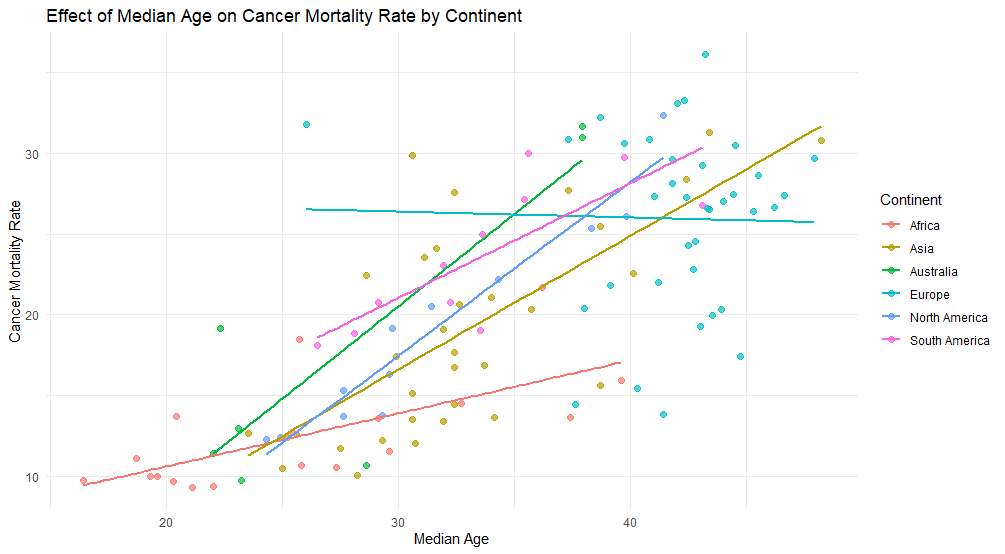
*נגדיר את להיות מידת זיהום האוויר כאשר קבוצת הבסיס היא minor, - medium, - high*

**2.4 הגדרה והוספה של משתני אינטראקציה:**

**2.4.1 משתנה אינטראקציה בין המשתנה Continent לבין המשתנה Median Age:**

הגיל החציוני במדינה עשוי להשפיע על תמותה מסרטן באופן שונה בין יבשות בשל הבדלים דמוגרפיים, תרבותיים וגישה לשירותי בריאות. ההשערה שלנו היא שבמדינות מפותחות (שנמצאות ביבשות מפותחות) עם גיל חציוני גבוה יותר, תמותת סרטן עשויה להיות גבוהה יותר בשל אוכלוסייה מבוגרת ואבחון מוקדם, בעוד שבמדינות מתפתחות, גיל חציוני נמוך יותר משקף דמוגרפיה צעירה שבה סרטן הוא פחות גורם מוות משמעותי. כפי שנראה בגרף, הקשרים בין הגיל החציוני לשיעור התמותה מסרטן משתנים בין היבשות, דבר שמדגיש את החשיבות של הוספת משתנה האינטראקציה Median Age \* Continent למודל. באפריקה, הקשר חיובי אך מתון, מה שעשוי לשקף את ההשפעה של מדינות עם גישה מוגבלת לשירותי בריאות, אך תיעוד טוב יותר במדינות עם גיל חציוני גבוה יחסית. באוסטרליה, הקשר חיובי וחזק מאוד, מה שמדגיש את השפעת האוכלוסייה המזדקנת והאבחון המוקדם ביבשת מפותחת זו. מנגד, באירופה, הקשר שלילי ומתון, מה שמראה הבדל ייחודי שדורש הבנה מעמיקה יותר של הגורמים התורמים, כמו מערכת בריאות מפותחת שייתכן שמפחיתה תמותה או מקטינה את השפעת הגיל. על כן הגרף מדגיש שהוספת משתנה האינטראקציה למודל תאפשר להבין טוב יותר את ההבדלים בין יבשות ותתרום להסבר מדויק ומפורט יותר של שיעור התמותה מסרטן. כמו כן, ניתן לראות שערך ה-p-value קטן מ-0.05, מה שמחזק את ההצדקה להכללת משתנה האינטראקציה במודל, כיוון שהוא תורם באופן מובהק להסבר השונות בשיעור התמותה מסרטן.



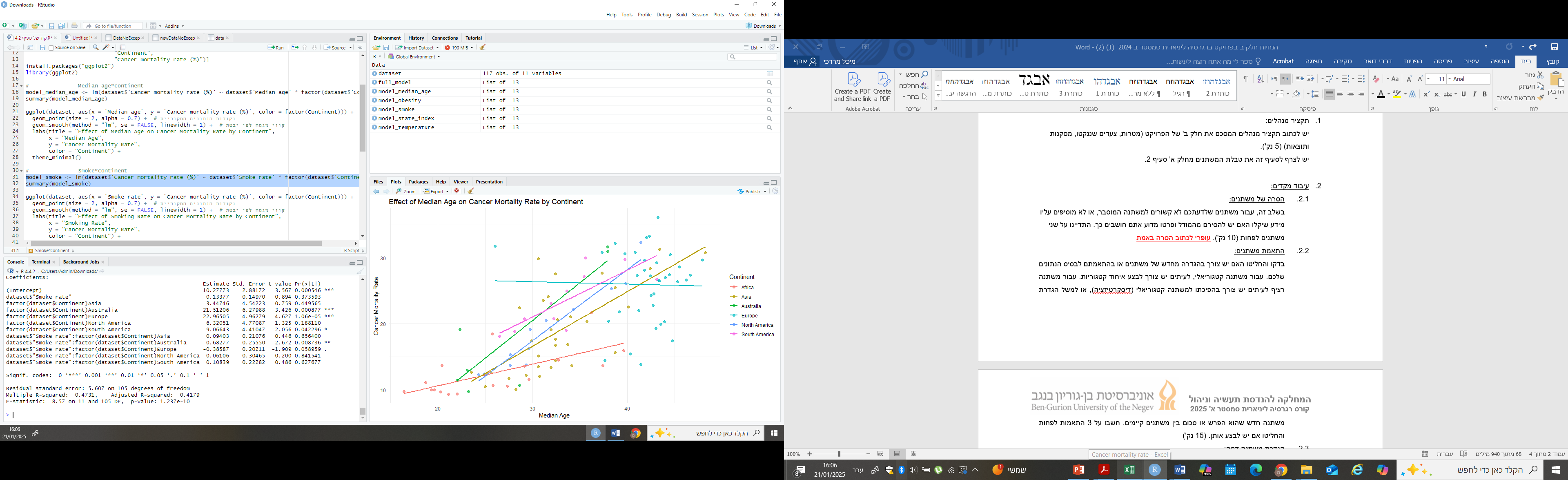


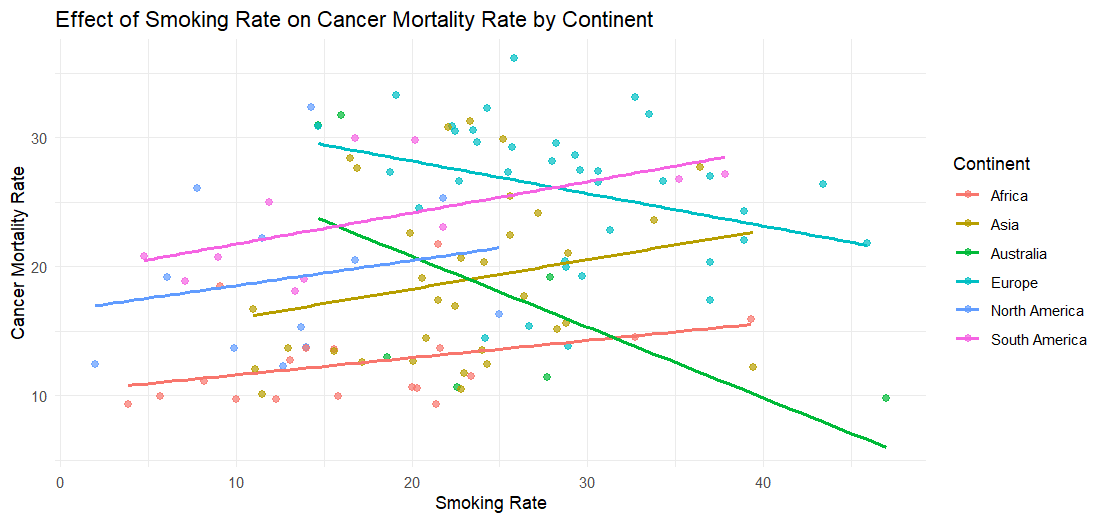
**2.4.2 משתנה אינטראקציה בין המשתנה הקטגוריאלי Continent לבין המשתנה Smoke rate:**

שיעור העישון הוא גורם סיכון ידוע לסרטן ולמחלות רבות נוספות, ולכן סביר להניח שהוא ישפיע על שיעורי התמותה מסרטן. עם זאת, ההשפעה עשויה להשתנות בין יבשות בשל הבדלים תרבותיים, בריאותיים ותשתיתיים. לדוגמה, באזורים שבהם שיעורי העישון גבוהים יותר ומערכות הבריאות פחות מפותחות, ניתן לצפות לקשר חזק יותר בין עישון לתמותה מסרטן. לעומת זאת, באזורים מפותחים עם גישה רחבה לטיפולים רפואיים ומודעות גבוהה לסיכוני עישון, הקשר עשוי להיות מתון או שונה בכיוונו.

כפי שניתן לראות בגרף, אכן קיים הבדל משמעותי בין היבשות. לדוגמה, ביבשות מפותחות כמו אוסטרליה (ירוק) ואירופה (טורקיז), ככל ששיעור המעשנים עולה, שיעור התמותה מסרטן דווקא יורד. קשר שלילי זה עשוי לנבוע ממערכות הבריאות המפותחות באזורים אלו, אשר מצליחות למתן את ההשפעה של עישון על תמותה. לעומת זאת, ביבשות פחות מפותחות כמו אפריקה (אדום), דרום אמריקה (ורוד), ואסיה (צהוב), נראה קשר חיובי בין עישון לתמותה מסרטן, מה שתואם את השערתנו. סביר להניח שבמדינות אלו מערכות הבריאות פחות מפותחות ולכן מתקשות להתמודד עם סיבוכי העישון ותוצאותיו. יבשת חריגה ביחס להשערתנו היא צפון אמריקה (תכלת), שבה הקשר בין עישון לתמותה חיובי אך מתון. ייתכן שהפער נובע ממאפיינים ייחודיים של המדגם שנבחר או מגורמים נוספים שמשפיעים על הקשר.

לסיכום, הגרף מדגיש את התרומה של משתנה האינטראקציה Smoke rate \* Continent להבנת ההשפעה המשתנה של עישון על תמותה מסרטן. הוא ממחיש כיצד ההשפעה של גורם סיכון ידוע עשויה להיות תלויה בהקשר הגאוגרפי, הכלכלי והבריאותי של האוכלוסייה. בנוסף, ערך ה-p-value הנמוך מ-0.05 מחזק את ההצדקה להוסיף את משתנה האינטראקציה למודל, שכן הוא מצביע על תרומה מובהקת להסבר השונות בשיעור התמותה מסרטן.





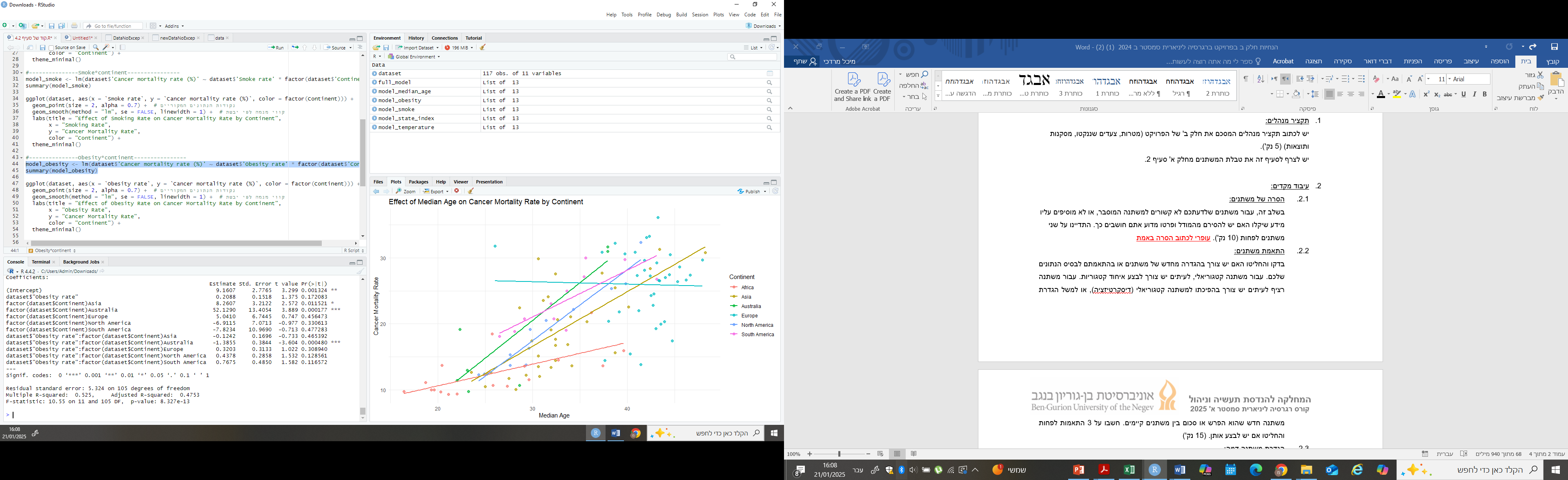
**4.2.3 משתנה אינטראקציה בין המשתנה Continent לבין המשתנה המסביר Obesity:**

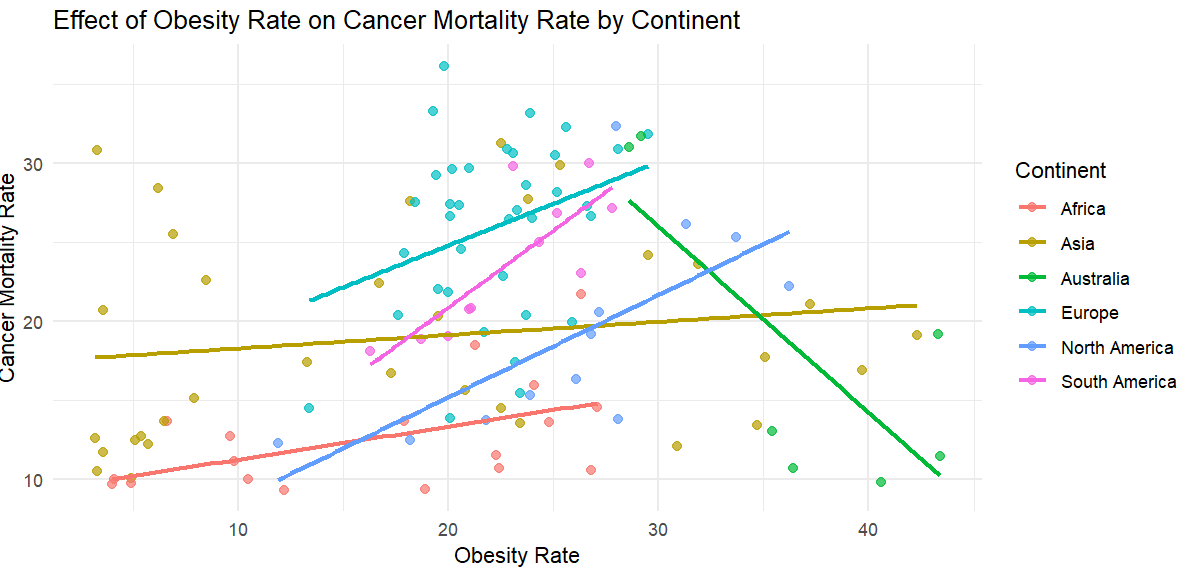
שיעור ההשמנה הוא גורם סיכון ידוע לסרטן ולמחלות כרוניות רבות נוספות, ולכן סביר להניח שהוא ישפיע על שיעורי התמותה מסרטן. עם זאת, ההשפעה עשויה להשתנות בין יבשות בשל הבדלים תרבותיים ורפואיים. לדוגמה, באזורים שבהם שיעורי ההשמנה גבוהים ומערכות הבריאות מתקשות לספק טיפולים מונעים, ניתן לצפות לקשר חיובי וחזק יותר בין השמנה לתמותה מסרטן. לעומת זאת, באזורים מפותחים עם מערכות בריאות מתקדמות ותוכניות מניעה הממוקדות בטיפול בהשמנה, הקשר עשוי להיות מתון, אפסי, או אף שלילי. כמו כן, ייתכן שבמדינות פחות מפותחות קיימת בעיה של חוסר אבחון למחלת הסרטן, כך שמקרים רבים של סרטן אינם מזוהים כלל. במקרים כאלה, ייתכן שנראה דווקא שיעורי תמותה נמוכים יותר מסרטן, לא בשל בריאות משופרת, אלא כתוצאה מחוסר יכולת לאבחן את המחלה ולהתחקות אחר השלכותיה.

בהתאם לגרף, ניתן לראות כי באוסטרליה קיימת מגמה שלילית בקשר בין המשתנים. מגמה זו עשויה לנבוע מכך שבאוסטרליה קיימת מערכת בריאות מתקדמת אשר מאבחנת את מרבית החולים במחלת הסרטן ומספקת להם טיפול יעיל שמונע את השפעות ההשמנה על תמותה. יכולת זו לשלב בין אבחון מוקדם לטיפול מתאים עשויה להסביר את השוני באוסטרליה ביחס לשאר היבשות. לעומת זאת, ביבשות פחות מפותחות כמו אסיה ואפריקה, ניתן להניח כי חלק גדול מהאוכלוסייה החולה בסרטן אינו מאובחן כלל, מה שעשוי להסביר מדוע נראה בגרף קשר מתון בלבד בין המשתנים. ייתכן כי היעדר תשתיות בריאותיות, גישה מוגבלת לשירותי רפואה, ומודעות נמוכה למחלה מובילים לכך שמקרי סרטן רבים אינם מדווחים, והמספרים הנמוכים בתמותה אינם משקפים את המציאות בפועל.

בשאר היבשות, כגון צפון אמריקה, דרום אמריקה, ואפילו אירופה, המגמה עולה וניתן לראות קשר חיובי בין המשתנים.השערתנו לכך היא שבמדינות אלו, מערכות הבריאות אמנם מצליחות לאבחן מספר רב של חולי סרטן, אך אינן תמיד מצליחות לספק את הטיפול הנכון או היעיל, במיוחד עבור אוכלוסיות שסובלות גם מהשמנה. הקשר החזק עשוי להצביע על כך שההשמנה מגבירה את הסיכון לתמותה מסרטן כאשר אין התמודדות מספקת עם גורמי הסיכון או עם המחלה עצמה.

לסיכום, הגרף מדגיש את השונות בין היבשות בהתמודדות עם הקשר בין השמנה לתמותת סרטן, ומראה את החשיבות של שימוש במשתנה אינטראקציה בין שיעור ההשמנה לבין היבשת. בנוסף, העובדה שערך ה-p-value קטן מ-0.05 מעידה על תרומה מובהקת של משתנה האינטראקציה למודל, ומצדיקה את הכללתו כחלק מההסבר לשונות בשיעור התמותה מסרטן.



**

*משתני האינטראקציה:*

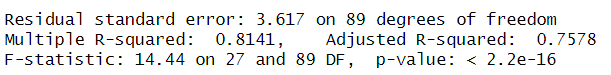
*נגדיר את להיות יבשת כאשר - אירופה, - אפריקה, - אוסטרליה, - דרום אמריקה, - צפון אמריקה. בנוסף, נגדיר את קבוצת הבסיס להיות אסיה ובהתאם אליה נבנה את משתני האינטראקציה:*

*משתני אינטראקציה בין שיעור העישון ליבשת:*

*משתני האינטראקציה בין השמנת יתר ליבשת:*

*משתני האינטראקציה בין גיל חציוני ליבשת:*

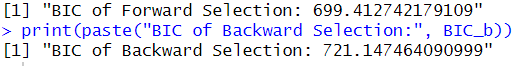
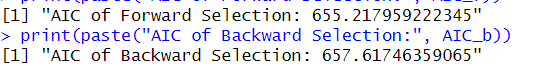
*המודל המלא:*

*******מדדי הביצוע:*

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

**3.1 בחירת משתני המודל:**כפי שלמדנו, ישנם 3 אלגוריתמים עיקריים נבחר 2 מהם:

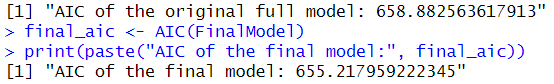
* Forward selection - נתחיל עם מודל ללא משתנים מסבירים, וננסה להכניס אליו משתנים (אחד-אחד). בכל צעד נבדוק איזה משתנה ישפר את המודל הכי הרבה, ואם השינוי מובהק נכניס אותו ונמשיך. אם אף הוספה לא תביא לשיפור מובהק (על פי מבחן F חלקי, למשל), או שנגמרו המשתנים - נעצור.
* Backward elimination - נתחיל עם מודל שמכיל את כל המשתנים. בכל צעד נבדוק איך תשפיע הוצאה של כל אחד מהמשתנים שבמודל. נוציא את המשתנה שהוצאה שלו תשפר את המודל הכי הרבה, במידה שהשיפור מובהק. כשאין עוד הוצאה שתביא לשיפור מובהק, נעצור.

****לאחר הרצה, בחרנו לבדוק קריטריון AIC, משמש להערכת איכות מודלים סטטיסטיים, מטרתו העיקרית היא למדוד כמה טוב מודל מסביר את הנתונים תוך כדי "ענישה" על מורכבות. בנוסף, התעניינו גם בקריטריון BIC. בפלט קיבלנו שמדד AIC קטן יותר וגם מדד BIC קטן יותר בForward מאשר בbackward, ולכן נעדיף את המודל שנוצר מהאלגוריתם זה.

**תמונה שמכילה טקסט, גופן, צילום מסך, קו

התיאור נוצר באופן אוטומטי**ה-Final Model שנבחר יהיה זהה למודל הסופי שיצא מ- Forward selection.

**השוואה בין קריטריון** **BIC בין המודל המקורי לבין המודל הסופי ->**

* ****בדיקה: אכן קריטריון ה-BIC השתפר (הצלחנו למזער).

**השוואה בין קריטריון** **AIC בין המודל המקורי לבין המודל הסופי ->**

* בדיקה: אכן קריטריון ה-AIC השתפר (הצלחנו למזער).

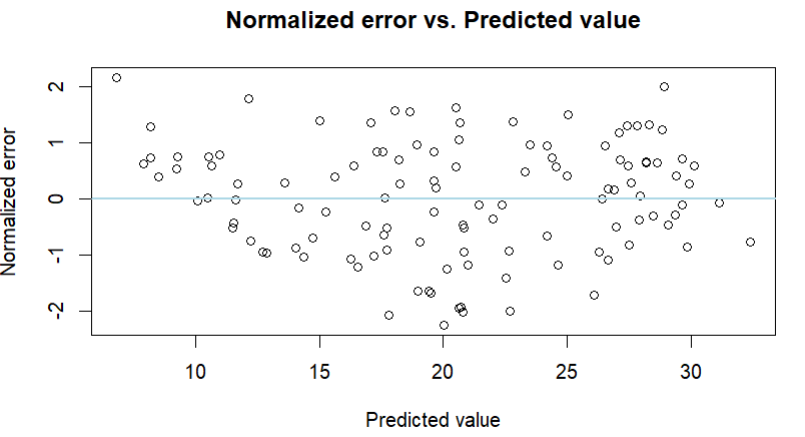
**השוואה בין Full Model המקורי לבין ה-Final Model :**

Full Model:

Final Model:

**3.2 בדיקת ויזואלית של הנחות המודל:**

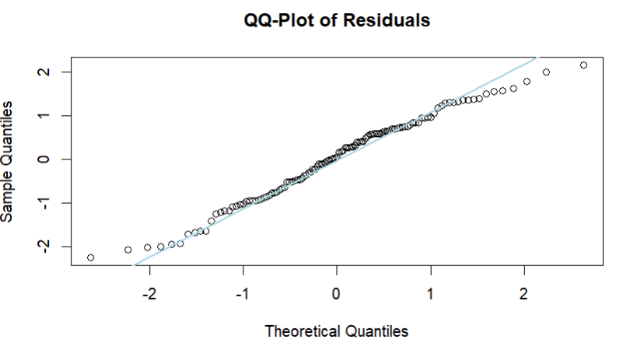
נרצה לבדוק אם הנחות הרגרסיה מתקיימות במודל שבנינו בסעיף 3.1. על מנת לבדוק את הנחת הלינאריות והנחת שוויון השונויות נבנה תרשים פיזור אשר מציג את השאריות המתוקננות אל מול התחזיות במודל.



הנחת שוויון שונויות:   
על מנת להניח שוויון שונויות, אנו מצפים לראות פיזור אחיד של השגיאות לאורך ציר התחזיות (ציר X). בבחינה ויזואלית של התרשים. ניתן להבחין כי הפיזור נראה אחיד למדי לאורך כל תחום התחזיות, ללא דפוס ברור של התרחבות או צמצום הפיזור (דמוי משפך). פיזור זה מעיד כי ניתן להניח שהנחת שוויון השונויות מתקיימת במודל.

הנחת הלינאריות:   
על מנת להניח לינאריות, אנו מצפים לראות פיזור אקראי של השגיאות סביב קו האפס לאורך ציר התחזיות (X) ללא דפוסים ברורים כמו עקומה או פרבולה, המעידים על קשר לא לינארי בין המשתנים המסבירים למשתנה התלוי. בבחינה ויזואלית של התרשים, נראה כי הפיזור של השגיאות אינו מראה דפוס עקבי או מגמה ברורה, מה שמחזק את ההנחה כי הקשר בין המשתנים במודל הוא לינארי. יחד עם זאת, מכיוון שמדובר בבדיקה ויזואלית, ייתכן שישנם קשרים לא לינאריים עדינים שאינם נראים לעין בתרשים זה. על כן, ניתן לשער כי הנחת הלינאריות מתקיימת.

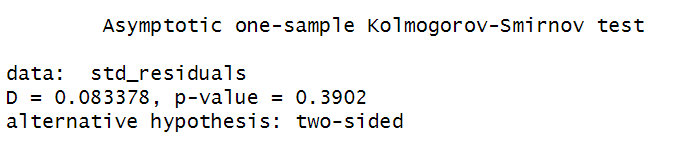
הנחת הנורמאליות:   
על מנת לבדוק את הנחת הנורמליות נבנה שני תרשימים QQ plot והיסטוגרמה של השגיאות המנורמלות

תמונה שמכילה תרשים, קו, עלילה, מדרון

התיאור נוצר באופן אוטומטי

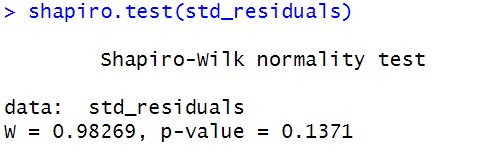
על מנת להניח שהשגיאות מתפלגות נורמלית, אנו מצפים לראות נקודות ב-QQ-Plot שעוקבות אחרי קו ישר, וההיסטוגרמה בצורת פעמון סימטרית. מהגרפים, ניתן לראות כי ב-QQ-Plot הנקודות עוקבות אחרי הקו אך לא בצורה מדויקת, עם סטיות בקצוות. ההיסטוגרמה מציגה מבנה כללי הדומה להתפלגות נורמלית, אך עם סימנים לסטיות.  
ניתן לשער שהנחת ההתפלגות הנורמלית מתקיימת באופן חלקי, עם חריגות מסוימות. על כן, מומלץ לבצע בדיקה מעמיקה יותר על מנת להחליט באופן סופי.

**3.3 בדיקה פורמלית של הנחות המודל:**פורמליות הנחת הנורמאליות:על מנת לבדוק באופן פורמלי את הנחת הנורמאליות במודל נבצע שני מבחנים KS וSW כך נדע האם לאשש או להפריח את ההשערות שהעלינו מהבדיקה הוויזואלית בסעיף הקודם:

****

**מבחן KS**

*בהתבסס על מבחן קולמוגורוב-סמירנוב (KS) שבו p-value = 0.3902 , ניתן לראות כי ערך ה-p גדול מ-0.05. מכיוון שאין עדות מספקת לדחות את השערת האפס, ניתן לומר כי ברמת ביטחון של 95%, ההתפלגות של השגיאות המתוקננות קרובה לנורמליות.*

******

**מבחן SW**

*בהתבסס על מבחן שפרו-ווילק (SW) שבו p-value = 0.1371 , ניתן לראות כי ערך ה-p גדול מ-0.05. מכיוון שאין עדות מספקת לדחות את השערת האפס, ניתן לומר כי ברמת ביטחון של 95%, ההתפלגות של השגיאות המתוקננות קרובה לנורמליות.*

*תמונה שמכילה טקסט, גופן, צילום מסך, לבן

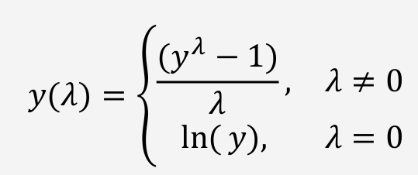
התיאור נוצר באופן אוטומטיפורמאליות הנחת הלינאריות:   
על מנת לבדוק את הנחת הלינאריות במודל נבצע את מבחן Chow, כך נדע האם לאשש או להפריח את ההנחות הוויזואליות מהסעיף הקודם:   
  
  
  
  
  
  
  
  
  
  
  
בהתבסס על מבחן M-fluctuation, שבו p-value = 0.005949 ניתן לראות כי ערך ה-p קטן מ-0.05. מכיוון שיש עדות חזקה לדחיית השערת האפס, ניתן להסיק ברמת ביטחון של 95% כי מבנה הקשרים במודל אינו לינארי. מכאן ניתן להסיק כי הנחת הלינאריות אינה מתקיימת במודל זה.*

**מבחן Chow**

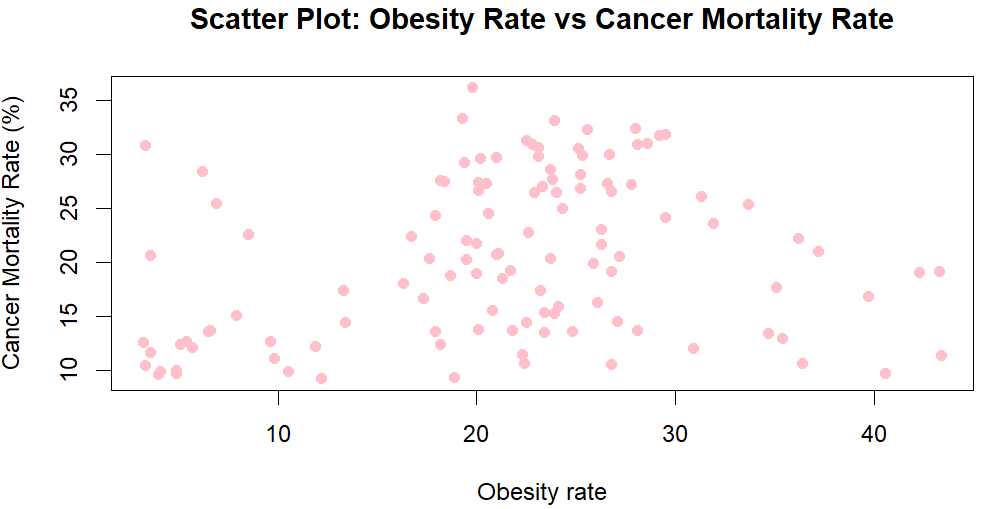
*תמונה שמכילה קו, תרשים, טקסט, עלילה

התיאור נוצר באופן אוטומטי****4. שיפור המודל:***

*במטרה לשפר את המודל, ביצענו מספר שלבים לבחינה והוספה של טרנספורמציות למשתנה המוסבר ולמשתנים המסבירים. בנוסף, שקלנו הוספת משתנים חדשים ובחינה מחדש של משתנים שהושמטו.*

*בשלב הראשון, בחנו באיזו טרנספורמציה נשתמש על המשתנה המוסבר באמצעות מבחן Box-Cox. המבחן הראה כי 0 נמצא בתוך הרווח סמך ולכן נבחר את הλ להיות שווה ל0. בהתאם, הפעלנו טרנספורמציית לוג על המשתנה המוסבר.*

*בשלב הבא, חיפשנו איזו טרנספורמציה תוכל לשפר את עמידת המודל בהנחת הלינאריות. באחת הבדיקות התבוננו בגרף של השמנת יתר ושיעור התמותה מסרטן:*

**

*בגרף ניתן לראות פיזור לא לינארי של הערכים עם שונות גדולה בין ערכים שונים. הנחנו שטרנספורמציית שורש תקטין את השפעת הערכים הקיצוניים, תמתן את השונות, ותסייע ליצור קשר לינארי יותר בין המשתנים ולכן עשויה לשפר את הלינאריות של המודל.*

*לאחר בדיקת מגוון טרנספורמציות אפשריות, במטרה לשפר את ערכי המדדים ואת העמידה בהנחות מודל הרגרסיה. המידול שהגענו אליו הוא המידול הבא:*

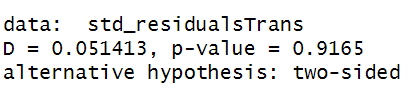
*לאחר ביצוע הטרנספורמציות, ניכר כי חל שיפור משמעותי במודל מבחינת מדד AIC כלומר הייתה ירידה משמעותית בערכי המדד, כמו כן היה שיפור גם במדד כלומר הייתה עליה בערכי המדד.*

|  |  |  |
| --- | --- | --- |
| *מדד* | *לפני טרנספורמציות* | *אחרי טרנספורמציות* |
| *AIC* | *655.2179* | *-40.3907* |
|  | *0.7442* | *0.7547* |

*כעת, המודל משופר ועומד בהנחות על סמך המבחנים שנבדקו:*

מבחן CHOW -עומד בהנחת הלינאריות

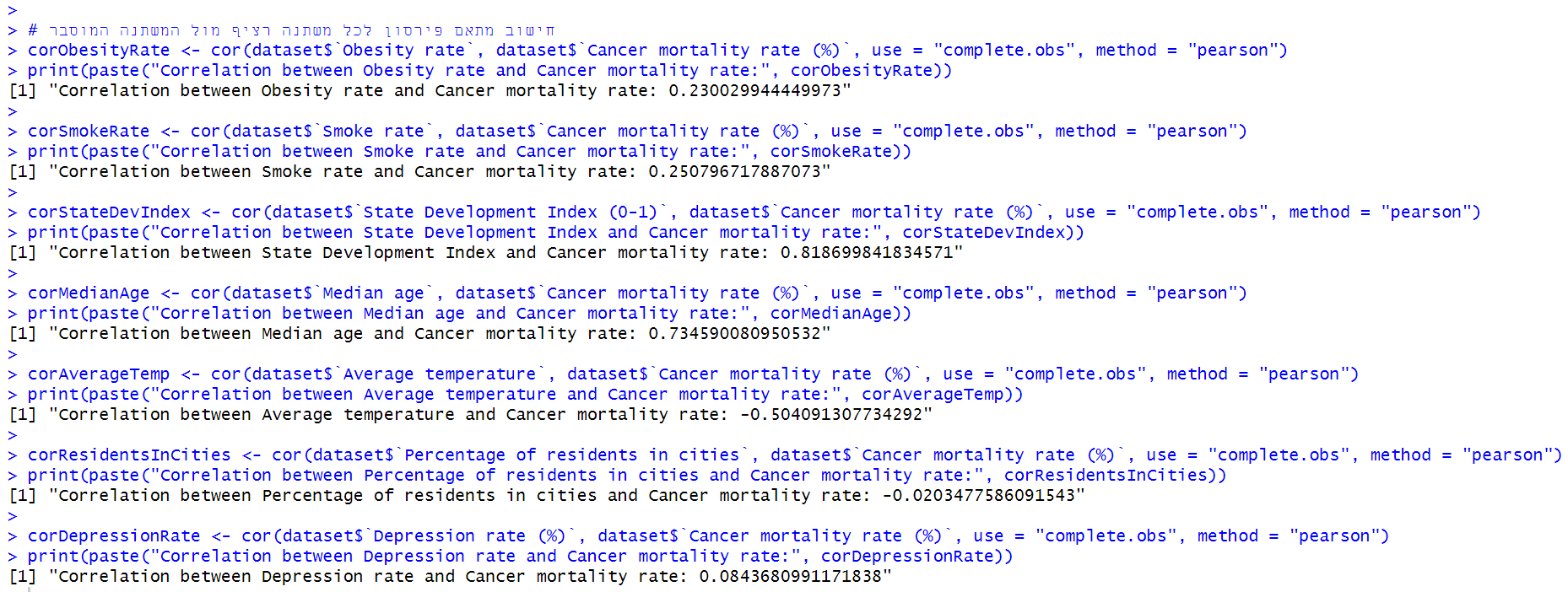
מבחן KS -עומד בהנחת הנורמליות

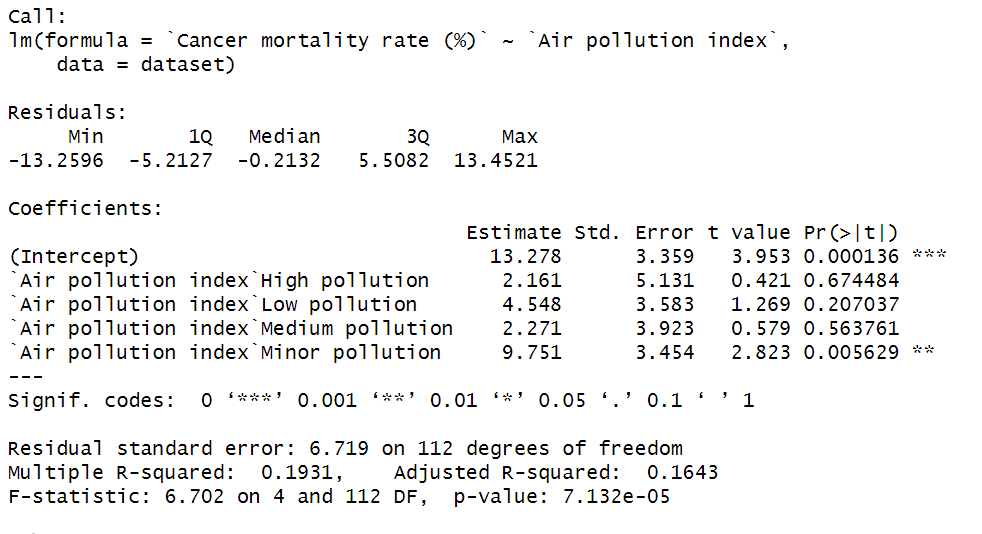
*תמונה שמכילה טקסט, גופן, לבן, טיפוגרפיה

התיאור נוצר באופן אוטומטי*

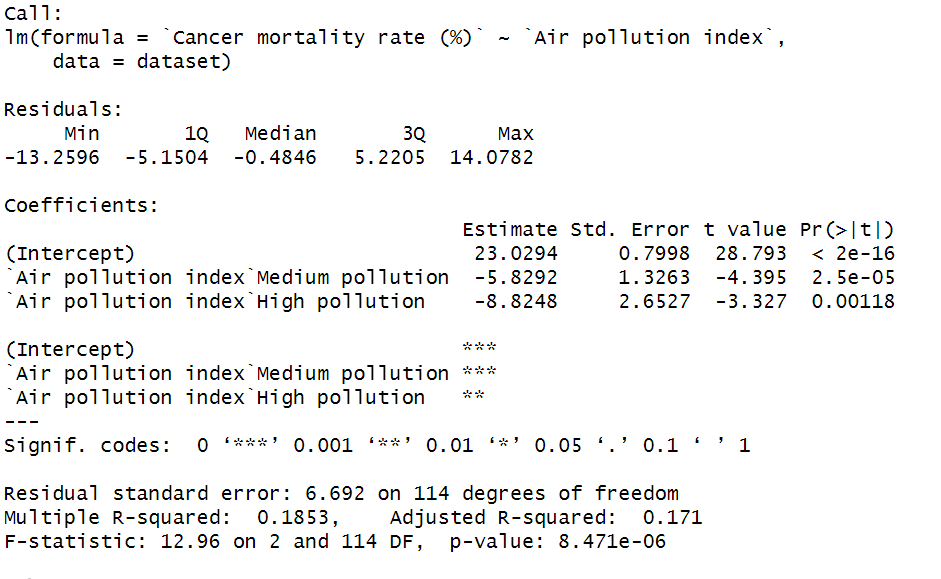
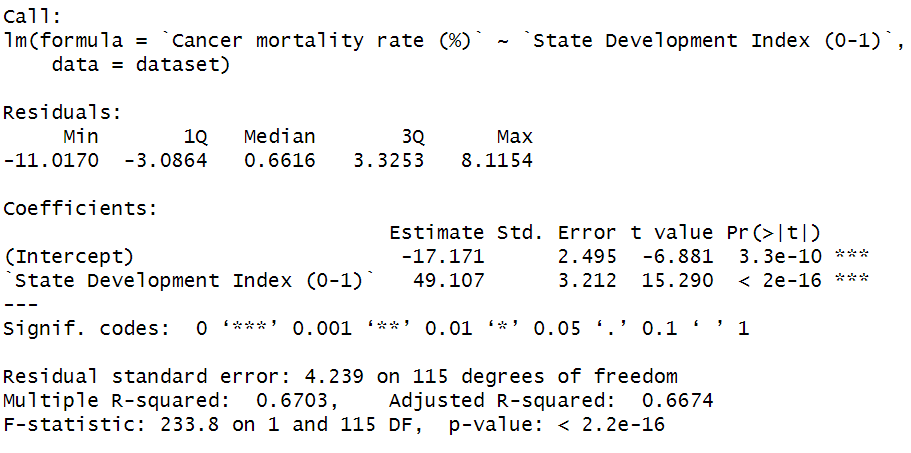
מבחן SW -עומד בהנחת הנורמליות

*נספחים:*

*סעיף 2.1*

**

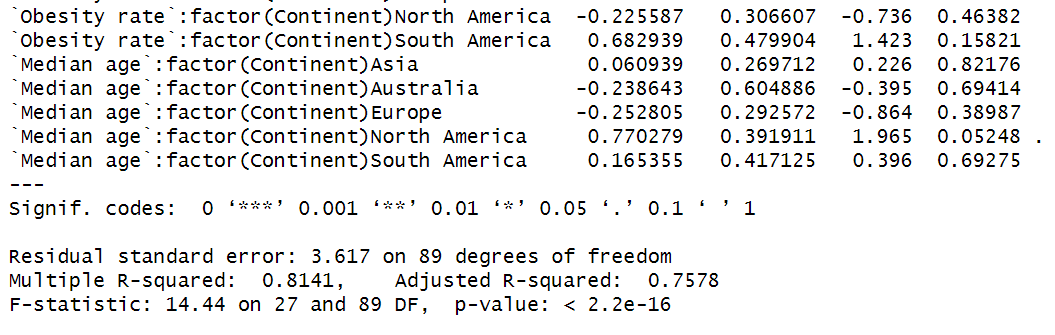
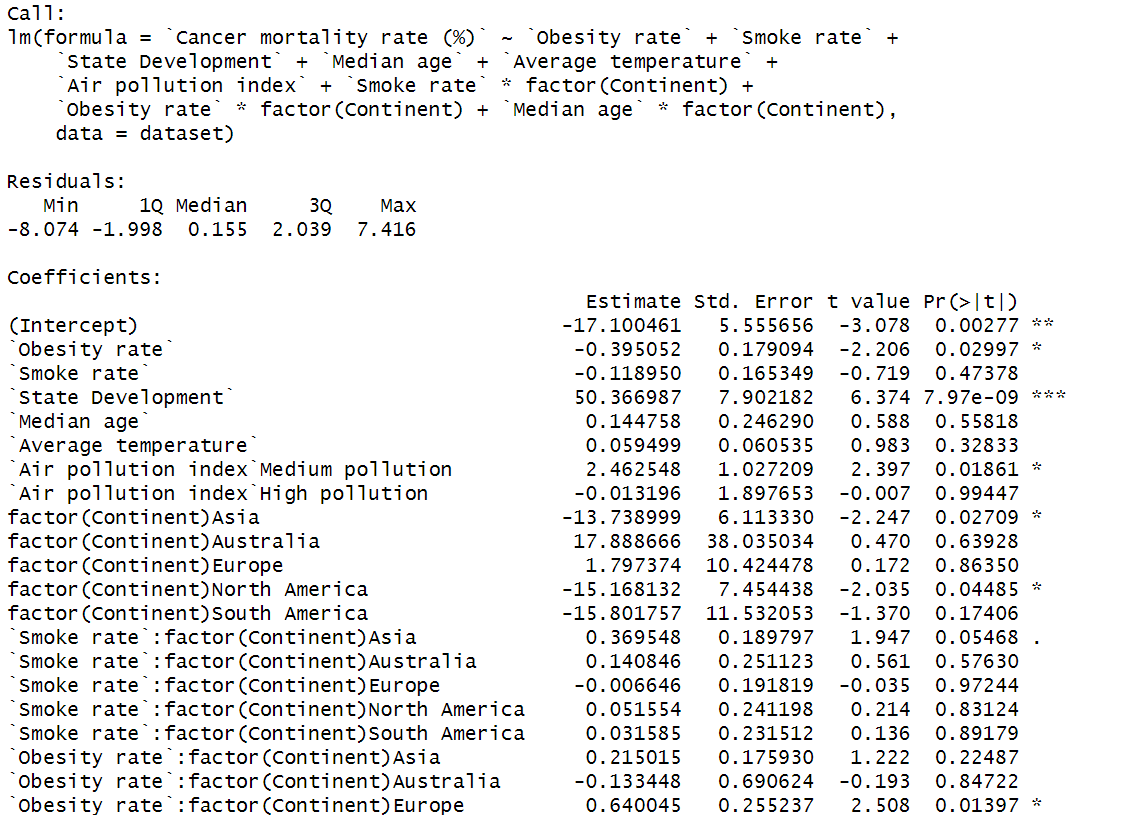
*סעיף 2.2*

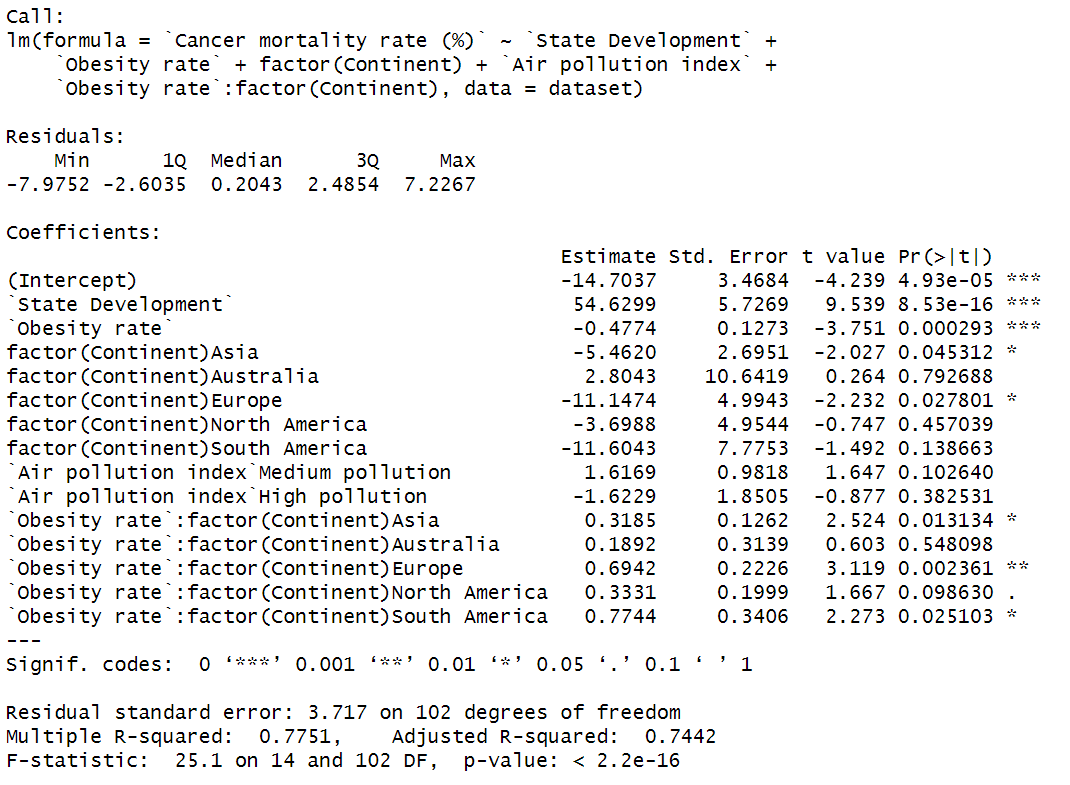
*  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
*

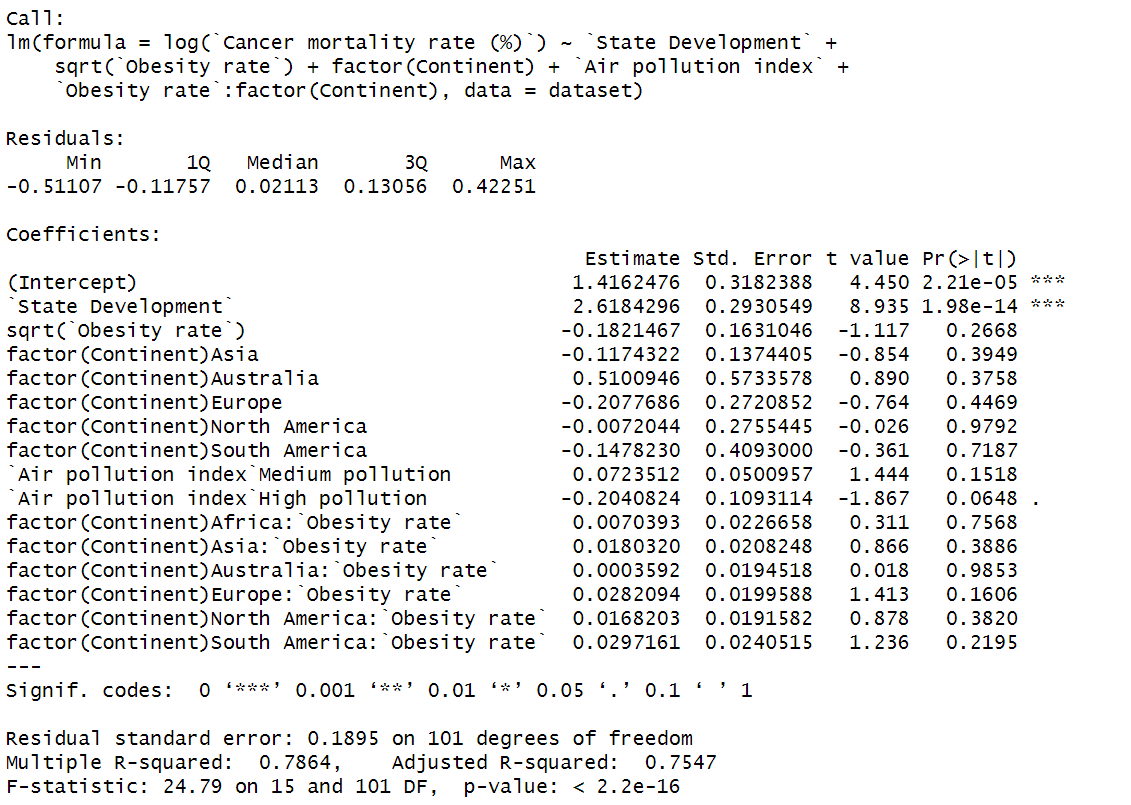
*תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטי*

*סעיף 2.4*



*  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
סעיף 3.1*

*סעיף 4*