



Department of
Mathematics

היחידה
למתמטיקה

**פרויקט מסכם לתואר בוגר במדעים (B.Sc)
במתמטיקה שימושית**

**A comparison between linear and quantile
regression in estimation of credit rating models**

Irena Machlev

**השוואה בין הרגרסיה הליניארית הקלסית לבין הרגרסיה
האחוזונית באמידת מודלים לדירוג אשראי**

אירנה מחלב

**השוואה בין הרגרסיה הליניארית הקלסית לבין הרגרסיה
האחוזונית באמידת מודלים לדירוג אשראי**

אירנה מחלב

**A comparison between linear and quantile
regression in estimation of credit rating models**

Irena Machlev

Advisor:
Mr. Levi Nir

מנחה:
מר לוי ניר

Karmiel

2016

כרמיאל

תקציר:

דירוג אשראי הינו כלי אשר פותח על מנת לאמוד את יכולת הלקוח להחזיר את הלוואתו בזמן. הבנק, מצד אחד, אינו רוצה להפסיד מכספו על ידי כך שייתן ריבית נמוכה מדי על סכום ההלוואה ומצד שני נמנע מלתת ריבית גבוהה מדי אשר הלקוח לא יוכל לעמוד בה, מה שיקטין את יכולת ההחזר שלו לבנק. לכן, צריך הבנק לתמרון בין כל הפרטים הידועים לו על הלקוח ועל השוק בכדי להציע את הריבית האופטימלית המתאימה ללווה ספציפי.

לשם כך פותחו מודלים רבים ושונים אשר לוקחים בחשבון נתונים על הלקוח ומנסים לגלות את יכולת ההחזר שלו על פי מידע קודם ולחזות את התנהגותו בעתיד.

בעבודה זו התמקדנו בשתי שיטות אקונומטריות לאמידת הריבית האופטימלית שהבנק יכול לתת לחברה אשר מבקשת הלוואה: שיטת הרגרסיה הליניארית הרגילה המשתמשת במזעור סכום ריבועי השגיאות ושיטת הרגרסיה האחוזונית, אשר פותחה בשנים האחרונות והוכחה כיעילה מאוד במקרים רבים.

נציג את שתי השיטות, נסקור ספרות עכשווית על מנת לבדוק את יתרונותיה של שיטת הרגרסיה האחוזונית ונבדוק את יעילות השיטות הנ"ל במציאת הריבית האופטימלית שיש לבנק להציע לחברה ספציפית.

לבסוף, ננסה להסיק באיזה סוג רגרסיה כדאי לבנק להשתמש על מנת לאמוד את הריבית האופטימלית שיש על הבנק להציע ללקוח.

1.	הקדמה.....	5
2.	סקירת ספרות.....	6
3.	מתודולוגיה.....	11
4.	סימולציות.....	20
5.	תוצאות ומסקנות.....	22
6.	סיכום ומחקרי המשך.....	25
7.	רשימה בבליוגרפית.....	27
	נספחים.....	28

הקדמה:

בשיטת מזעור סכום ריבועי השגיאות של הרגרסיה הליניארית הקלאסית, פונקצית הממוצע אשר מתארת כיצד משתנה הממוצע של המשתנה המוסבר יחד עם וקטור המשתנים המסבירים, הוא בדרך כלל רוב המידע שאנו צריכים על מנת לאמוד את יחסי הגומלין בין המשתנה המוסבר למשתנים המסבירים. במקרה זה אנו חייבים להניח כי התפלגות ההפרעות זהה להתפלגות המשתנים המסבירים. ואם אכן כך, אנו יכולים להיות מרוצים מתוצאות המתקבלות מאמידה בעזרת מודל הממוצע המותנה אשר מציעה הרגרסיה הליניארית הרגילה, אך מה קורה כאשר אנו לא בטוחים לגבי ההתפלגות של המתשנים וההפרעות?

הרגרסיה האחוזונית, אשר הוצגה לראשונה על ידי באסט וקוהנקר בשנת 1978, עונה על השאלות בצורה ספציפית יותר ומתייחסת להתפלגות המשתנים על ידי חלוקתם לאחוזונים. למשל, משכורותו של עובד תוגדר להיות באחוזון ה- Θ של התפלגות המשכורות אם הוא מרוויח יותר מחלק Θ של העובדים הנבדקים ופחות מ- $(1-\Theta)$ מהעובדים. הרעיון הוא לאמוד את הפונקציות המותנות בכל אחוזון, על ידי מודלים בהם האחוזונים של ההתפלגות המותנה של המשתנה המוסבר מבוטאים על ידי פונקציות המשתנים המסבירים.

כפי שבעזרת הרגרסיה הליניארית הרגילה ניתן להציג את סכום הריבועים המזעריים של השגיאות, כך ברגרסיה האחוזונית ניתן להציג את סכום הריבועים המזעריים המשוקלל של השגיאות לכל אחוזון על פי התפלגותו.

מכיוון שמטרתנו לבדוק באיזו שיטה מתקבל חיזוי מדויק יותר של הריבית האופטימלית, בשיטת הרגרסיה הליניארית הרגילה או בשיטת הרגרסיה האחוזונית, נשווה בין הפרשי השגיאות המתקבלות בשתי השיטות על ידי הרצת מודל בתוכנת ה-MATLAB. המודל יורץ על 10,000 חברות, כאשר לכל חברה יהיו ידועים לבנק שלושה פרמטרים: ותק החברה, הכנסות החברה וסך החובות אשר יש לחברה.

בפרק **סקירת הספרות** נציג סקירה מתמצתת של ספרות בנושא הרגרסיה הליניארית הקלאסית ושל הרגרסיה האחוזונית ונראה את שימושיה העיקריים.

בפרק **המתודולוגיה** נציג את המודלים ואופן האמידה בכל אחד מהם, בנוסף נציג את בעיית האינדוגניות היכולה להיווצר והדרכים להתמודד איתה.

בפרק **הסימולציה** תבוצע סקירה על נתוני המדגם, אופן בניית תוכנית ה-MATLAB והתוצאות.

בפרק האחרון, פרק **הסיכום**, יוצגו המסקנות והמלצות למחקרי המשך.

סקירת ספרות:

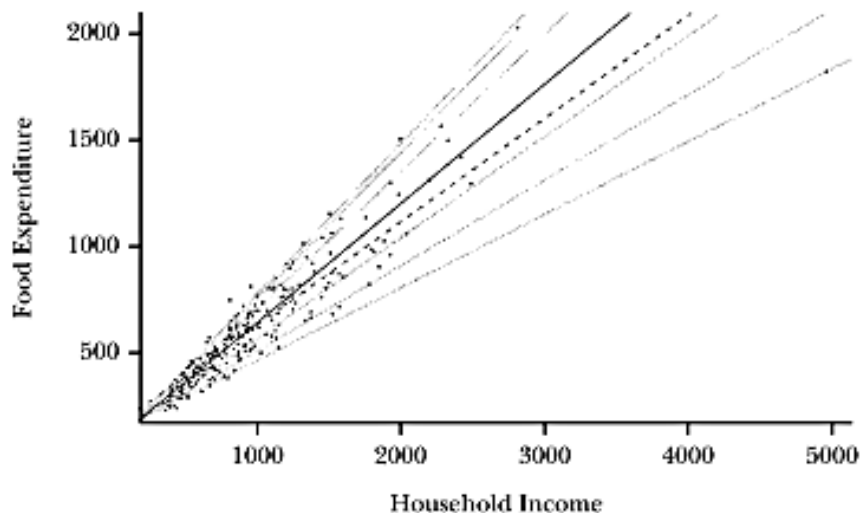
בעבודה זו התמקדנו בבדיקת יעילות שיטת הרגרסיה האחוזונית במקרים מסויימים לעומת הרגרסיה הליניארית הקלסית.

בשנים האחרונות, לרגרסיה האחוזונית שימושים נרחבים בספרות המקצועית. נראה כי השימוש בסוג זה של רגרסיה, כאשר המשתנים מרובים, מתאר בצורה טובה יותר את המגמות והמסקנות ממוקדות יותר לפי כל אחוזון.

במאמרם, הציגו קוהנקר והאלוק את תוצאות עבודתו של אנגל¹(1857) אשר בדק את הקשר בין הוצאות משק הבית על מזון לבין ההכנסות שלו ב- 235 משפחות אירופאיות ממעמד הפועלים בעזרת רגרסיה ליניארית פשוטה. קוהנקר והאלוק רצו להציג את ההבדל במסקנות שניתן להסיק כאשר משתמשים ברגרסיה האחוזונית לעומת הרגילה.

בגרף הבא מוצגים שבעה קווי רגרסיה אחוזונית האומדים את הקשרים ביחס לשבעה אחוזונים [0.05,0.1,0.25,0.5,0.75,0.9,0.95].

קו הרגרסיה של החציון (האחוזון ה-50) מסומן בקו המודגש ואילו קו הרגרסיה המופק בשיטת הריבועים המזעריים מסומן בקו המקווקו.



¹ ממאמרם של קוהנקר והאלוק - Journal of Economic Perspectives—Volume 15, Number 4—Fall 2001— Pages 143–156.

לפי הפיזור ניתן לראות כי ככל שגדלות הכנסותיה של משפחה, כך עולות הוצאותיה על מזון. כמו כן, ההתפלגות המותנה של ההוצאות על מזון סוטה שמאלה, מה שמסביר את ההבדל בין האמד לרגרסיה האחוזונית לבין הרגרסיה הרגילה.

דבר נוסף אשר השפיע מאוד על קו הרגרסיה הרגילה הם שני ערכים קיצוניים ולא אופייניים של הכנסות נמוכות מאוד מחד והוצאות גבוהות על אוכל, מאידך.

נראה כי שיטת הרגרסיה הליניארית הרגילה מושפעת יותר מערכים קיצוניים ואינה משקפת לגמרי את מצבם של המשפחות העניות ביותר. מכיוון שהרגרסיה האחוזונית מתחשבת בהתפלגות של כל אחוזון, ניתן לראות כי התמונה שהיא מציגה ברורה יותר וכי שיטה כזו אמידה הרבה יותר לערכים קיצוניים ובו בעת לוקחת אותם בחשבון.

בדוגמה נוספת, ניתן לראות את תוצאות המחקר "השפעות על משקל התינוק בלידה" המוצגות במאמרה של סימונה דספה² (2007). המחקר בדק מהו משקל ההשפעות החיצוניות השונות על משקל הילוד בשתי שיטות: הרגרסיה הליניארית הרגילה והרגרסיה האחוזונית. בשיטה האחוזונית חולק משקל התינוקות שבמדגם לפי אחוזונים.

הטבלה הבאה מציגה את מקדמי הרגרסיה הרגילה והאחוזונית ומשווה ביניהם:

Characteristic	Linear Regression	Quantile Regression				
		5 th	10 th	50 th	90 th	95 th
Intercept	3224	2353	2608	3252	3856	4031
Married	161.1	227	171	149	141	165
Boy	115.9	28	84	121	142	142
Prenatal care	-227.0	-536	-418	-164	-111	-57
Smoke	-200.9	-255	-226	-190	-177	-199

קל לראות כי התוצאות אשר התקבלו בשימוש ברגרסיה האחוזונית מרכיבים תמונה בהירה יותר על השפעת המשתנים המסבירים. למשל, ניתן להסיק מהטבלה כי חוסר בטיפול טרום לידתי(מסומן באדום) משפיע בצורה שלילית רבה יותר על אחוזוני המשקל הנמוכים וכי מין הילוד משפיע הרבה יותר על משקל התינוק באחוזונים הגבוהים. החלוקה לאחוזונים מספקת הסבר מפורט יותר למשקלים השונים ומספקת כוח ניבוי מדויק יותר.

² Cornell University, StatNews #70: Quantile Regression November 2007 Updated 2012

כמו כן, במחקר: "מדידת פערי השכר בין המגזר הציבורי למגזר הפרטי בישראל"³ אשר נעזר בממצאים משימוש ברגרסיה אחוזונית, בדקו החוקרים וגילו כי השימוש ברגרסיה מסוג זה לעומת הרגרסיה הליניארית הקלאסית מוצדק, מכיוון שמסתבר כי הקשר בין התפלגות שכר העובדים למשתנים המסבירים מהווה נדבך חשוב באפיון הפרמיה לעובדי המגזר הציבורי ולא די באומדן הריבועים הפחותים.

על מנת להתגבר על חוסר הדיוקים ולקבל תוצאות נרחבות יותר, יש לאמוד את התפלגות הפרמיה גם לאורך האחוזונים השונים ולא רק בתוחלת, מה שמצדיק שימוש ברגרסיה אחוזונית במקרה זה.

דוגמא למחקר נוסף אשר השימוש ברגרסיה אחוזונית הביא לניבוי מדויק יותר של הצפות באיזור אנגליה מובאת במאמר: Estimation of predictive hydrological uncertainty using quantile regression⁴. למרות שנסיונות רבים נעשו באמידת כמות המשקעים וניבויים באזור המדובר, ישנם אי דיוקים וחוסר וודאות המלווים את התחזיות. כותבי המאמר מציעים להעריך את הקשר בין התחזית ההידרולוגית לבין סכום הפרשי השגיאות המתקבלות בתחזיות.

במקרה זה, השיטה אינה משתמשת באי הודאות של המשתנים הבלתי תלויים, אלא במקום, משתמשת באי הודאות האפקטיבית של תהליך החיזוי, אשר יכולה לנבוע מאי ודאות של כמה גורמים - כניסות ויציאות, מבנה המודל, או אי וודאות של פרמטרים במודל. לשם כך, חולקו ריבועי הפרשי השגיאות לאחוזונים והקשרים של השגיאה האחוזונית מוערכים במישור הגאוסיאני.

כאשר נבדקה התפלגות השגיאות בכל אחוזון בנפרד והמודל הורץ על תקופות שונות התקבלה תמונה ברורה מאוד לגבי האופן שבו ניתן לחזות את מועדי וסדרי גודל של ההצפות.

בעבודתו "changes in the U.S wage structure 1963-1987"⁵, משתמש פרופ' משה בוצ'ינסקי ברגרסיה האחוזונית בכדי לחקור בצורה מדוקדקת יותר את השפעת ההשכלה ונסיון העובדים על השינוי במשכורות בשנים המדוברות. השינוי במבנה השכר על פני תקופה זו הוא דרמטי, ומוסבר על ידי כמה סיבות שנבדקו בעבר: העלאת שכר המינימום, המודעות הגוברת לחוק שוויון העסקה בין שחורים ללבנים בארה"ב, הרחבת ההשכלה, טיב בתי הספר, הדרישה לעובדים משכילים יותר ועוד. אך בוצ'ינסקי רצה לדעת כיצד משפיעים רמת ההשכלה ונסיון העובד בנקודות שונות של התפלגות המשכורות במהלך השנים ובאיזו נקודה הייתה השפעה של איזה גורם גדולה יותר.

³ ירושלים, אפריל 2004

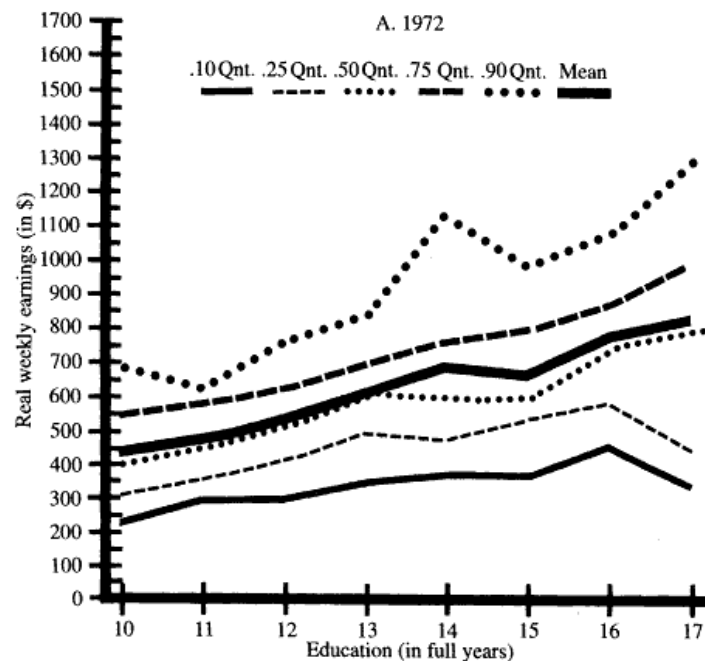
⁴ Weerts, A.H. · Winsemius, H.C. · Verkade, J.S., 2011

⁵ Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression (1994)

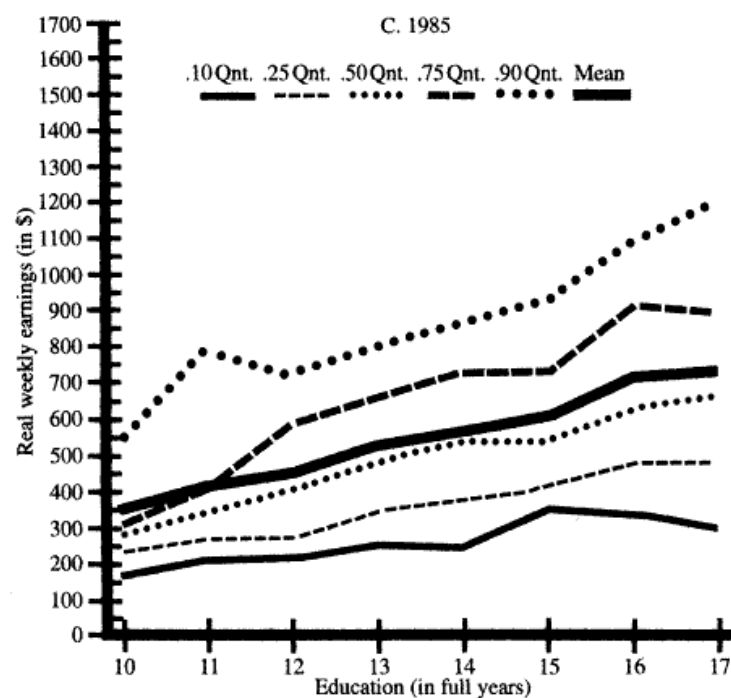
לשם כך, חילק החוקר את מדגם העובדים לפי קבוצות גיל ואת המשכורות השבועיות לחמישה אחוזונים, כאשר בכל אחוזון נבדקת רמת ההשכלה וכמות הנסיון של כל עובד.

בעזרת החלוקה לאחוזונים, התמקדות בהתפלגות המיוחסת לכל אחוזון בנפרד, הגיע החוקר לתוצאות מרחיקות לכת אשר אינן נראות בשימוש במודלים אקונומטריים אחרים.

נציג, לדוגמה את השפעת מספר שנות ההשכלה על שכרם של עובדים באותה קבוצת גיל לשנת 1972, בעלי נסיון של 15 שנים:



לעומת ההשפעה על שכרם של אותה קבוצה בשנת 1985:



כאשר נבדקו הדברים לאורך כל השנים 1963-1987 התגלה, למשל, כי התוצאות מציגות דפוסים דומים של צמיחה בשכר לאורך האחוזונים השונים אך שונים מהותית ברמות הצמיחה. כמו כן, השפעת רמת ההשכלה גבוהה יותר באחוזונים הגבוהים והשפעת רמת הנסיון גבוהה יותר באחוזונים הנמוכים, בעיקר בשנות השבעים המוקדמות. בכללי, השפעת רמת ההשכלה על השכר גבוהה יותר בקרב העובדים חסרי הנסיון באחוזונים הגבוהים ומשפיעה פחות על שכר העובדים הוותיקים יותר.

לעומת זאת, נסיון העובד מוערך כמשפיע יותר באחוזונים הנמוכים על העובדים החדשים ומשפיע פחות באחוזונים הגבוהים על העובדים הוותיקים. בנוסף, השפעת רמת הנסיון של עובד לא מנוסה על משכורתו קטנה עם השנים. אין ספק כי ממצאים אלו מציירים תמונה רחבה וברורה יותר לגבי השפעות המשתנים המסבירים על השינוי בהרכב השכר.

לסיכום, ניתן להגיד כי במקרים רבים הוכח השימוש ברגרסיה האחוזונית כיעיל יותר לתחזיות וחסין יותר לערכים קיצוניים. כמו כן, הרצת המודל על אחוזון נבחר בהתחשב בהתפלגות הנתונים באחוזון זה, נותנת הסברים רחבים יותר לתופעות השונות ועונה על שאלות ספציפיות יותר.

מתודולוגיה:

מטרת העבודה הינה אמידת הריבית האופטימלית עבור הבנק ומזעור פונקצית ההפסד שלו על ידי שימוש בשיטת סכום הריבועים המזעריים ברגרסיה הליניארית מרובת המשתנים והשוואתה לשיטת הרגרסיה האחוזונית.

נציג את שיטת הריבועים המזעריים, את הרגרסיה מרובת המשתנים ואת מודל הרגרסיה האחוזונית:

1. שיטת הריבועים המזעריים (OLS)

שיטת הריבועים המזעריים הינה שיטת אמידה אקונומטרית בסיסית, עבור מודלים לינאריים. שיטה זו מבטיחה אומדים בלתי מוטים, יעילים ועקיבים, בתנאי שמתקיימות 5 ההנחות הקלאסיות של מודל הרגרסיה.

5 הנחות המודל:

1. נכונות המודל התיאורטי – השפעתם של המשתנים המסבירים על המשתנה התלוי היא ליניארית.
 2. $E[\varepsilon_i] = 0$ - תוחלת ההפרעות האקראיות היא 0. מספיק לדרוש כי תוחלת ההפרעות האקראיות הינה קבועה, מניחים 0 מטעמי נוחות בלבד.
 3. $V[\varepsilon_i] = \sigma^2$ - שונות ההפרעות האקראיות הינה קבועה.
 4. $Cov[\varepsilon_i, \varepsilon_j] = 0, i \neq j$ - חוסר מתאם סדרתי, ההפרעות האקראיות של שני גורמים שונים אינן מתואמות.
 5. $Cov[x_i, \varepsilon_i] = 0$ - הנחת האקסוגניות של המשתנים המסבירים. המשתנים המסבירים לא מתואמים עם ההפרעה האקראית.
- יהי X המשתנה המסביר, Y המשתנה המוסבר ו- ε ההפרעה האקראית, נניח כי קיים קשר בין המשתנים:

$$y = \alpha + \beta x_i + \varepsilon_i$$

נרצה למצוא את האומדים ל- α ול- β . נסמנם ב- $\hat{\alpha}, \hat{\beta}$.

נמצא את האומד למשתנה β :

מכיוון שמטרתנו היא למצוא קו העובר בין התצפיות, הממזער את ריבועי המרחק (הטעות) בין כל תצפית אל הקו, נביא למינימום את הביטוי:

$$\min \sum (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2$$

$$1) \frac{\partial}{\partial \hat{\alpha}}: -2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$2) \frac{\partial}{\partial \hat{\beta}}: -2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0$$

נסמן את שגיאת האומד¹ כך: $u_i = y_i - \hat{\alpha} - \hat{\beta} x_i$

נראה כי מתנאי סדר ראשון נובע כי הרגרסיה עוברת דרך נקודת הממוצעים:

$$1) -2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\sum y_i = \sum \hat{\alpha} + \hat{\beta} \sum x_i$$

$$I \bar{y} = I \hat{\alpha} + \hat{\beta} I \bar{x} \rightarrow \bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$$

נעבור לתנאי השני ונציב את $\hat{\alpha}$:

$$2) -2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\sum (y_i - \bar{y} + \hat{\beta} \bar{x} - \hat{\beta} x_i) x_i = 0$$

$$\sum (y_i - \bar{y}) x_i = \hat{\beta} \sum (x_i - \bar{x}) x_i$$

מכיוון משתקיים: $\sum (y_i - \bar{y}) x_i = \sum (y_i - \bar{y})(x_i - \bar{x}) = \sum (x_i - \bar{x}) y_i$

האמד ל- β יהיה:

$$\hat{\beta} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})(x_i - \bar{x})}$$

לפי תכונות התוחלת ניתן להראות כי $\hat{\beta}$ הינו אומד חסר הטיה².

¹ שגיאת האומד – ההפרש בין ה-Y האמיתי לבין ל-Y שנאמד על ידי האומדים.

² אומד הינו חסר הטיה כאשר מתקיים: $E[\hat{\beta}] = \beta$

2. רגרסיה מרובת משתנים:

במקרה שלנו מדובר ברגרסיה עם כמה משתנים ולכן משוואת הרגרסיה תיראה כך :

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i$$

נרצה למזער את בעיית סכום ריבועי השגיאות אשר מבטאות את המרחק בין הערך האמיתי של המשתנים התלויים לבין ערכם החזוי :

$$\min_{\hat{\alpha}, \hat{\beta}, \hat{\gamma}} \sum (y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} z_i)^2$$

נגזור לפי $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$:

$$1) \frac{\partial}{\partial \hat{\alpha}}: -2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} z_i) = 0$$

$$2) \frac{\partial}{\partial \hat{\beta}}: -2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} z_i) x_i = 0$$

$$3) \frac{\partial}{\partial \hat{\gamma}}: -2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} z_i) z_i = 0$$

הראינו כי רגרסיה מסוג זה תעבור דרך נקודת הממוצעים, לכן :

$$\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x} + \hat{\gamma} \bar{z}$$

נציב את $\hat{\alpha}$ במשוואות 1,2,3 נעביר אגפים ונקבל :

$$2) \sum y_i - (\bar{y} - \hat{\beta} \bar{x} - \hat{\gamma} \bar{z}) - \hat{\beta} x_i - \hat{\gamma} z_i = 0$$

$$\sum (y_i - \bar{y}) x_i = \hat{\beta} \sum (x_i - \bar{x}) x_i + \hat{\gamma} \sum (z_i - \bar{z}) x_i$$

$$3) \sum (y_i - \bar{y}) z_i = \hat{\beta} \sum (x_i - \bar{x}) z_i + \hat{\gamma} \sum (z_i - \bar{z}) z_i$$

נבטא את $\hat{\beta}$ ו- $\hat{\gamma}$:

$$\hat{\beta} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x}) - \hat{\gamma} \sum (z_i - \bar{z})(x_i - \bar{x})}{\sum (x_i - \bar{x}) x_i}$$

$$\hat{\gamma} = \frac{\sum (y_i - \bar{y})(z_i - \bar{z}) - \hat{\beta} \sum (x_i - \bar{x})(z_i - \bar{z})}{\sum (z_i - \bar{z}) z_i}$$

נציב $\hat{\beta}$ ב- $\hat{\gamma}$ ונקבל:

$$\hat{\beta} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x}) \sum (z_i - \bar{z}) z_i - \sum (y_i - \bar{y})(z_i - \bar{z}) \sum (x_i - \bar{x})(z_i - \bar{z})}{\sum (z_i - \bar{z}) z_i - (\sum (x_i - \bar{x})(z_i - \bar{z}))^2}$$

באותה הצורה ניתן למצוא את האומד ל- $\hat{\gamma}$:

$$\hat{\gamma} = \frac{\sum (y_i - \bar{y})(z_i - \bar{z}) \sum (x_i - \bar{x}) x_i - \sum (y_i - \bar{y})(x_i - \bar{x}) \sum (x_i - \bar{x})(z_i - \bar{z})}{\sum (x_i - \bar{x}) x_i \sum (z_i - \bar{z}) z_i - (\sum (x_i - \bar{x})(z_i - \bar{z}))^2}$$

נציין, כי אם אין מתאם³ בין X ל-Z, ניתן להשתמש במודל הליניארי הפשוט כדי לאמוד את הקשר בין X ל-Y גם אם המודל הכולל את Z הוא הנכון. וזאת מכיוון שאם מתקיים:

$$\sum (x_i - \bar{x})(z_i - \bar{z}) = 0$$

נקבל אומד ל- β הזהה לאומד שהתקבל בחישוב הרגרסיה הפשוטה.

³ מתאם (קורלציה) - מדד סטטיסטי המעריך את העקביות ביחסים בין מספר משתנים, כלומר, האם ישנה עקביות בין שינוי במשתנה אחד לבין שינוי במשתנה השני.

3. רגרסיה אחוזונית:

שיטת הרגרסיה האחוזונית (quantile regression) הינה הרחבת שיטת האמידה הליניארית מרובת המשתנים.

הרגרסיה האחוזונית, אשר הוצגה לראשונה בשנת 1978 ע"י קואנקר ובאסט, שימושית מאוד בעיקר כאשר ישנם נתונים קיצוניים וכאשר התפלגות המשתנים Y המובעים על ידי מקדמי הרגרסיה, תלויים באחוזון. היא מספקת תמונה רחבה וברורה יותר של התפלגות המשתנה Y מכיוון שהיא מאפשרת אמידה לאורך תחום כל האחוזונים המותנים האפשריים של הפונקציה וניתן לקבל מידע על הקשר בין המשתנה התלוי לבלתי תלוי תוך התחשבות בצורת ההתפלגות של Y בכל אחוזון ואחוזון. לאומדני הרגרסיה האחוזונית יתרון הנובע, בין השאר, מהיכולת לקשר בין האומדנים לבין התפלגות המשתנה המסביר.

עבור כל משתנה רנדומלי תלוי Y , ניתן לומר כי הוא נמצא מעל אחוזון ה- Θ אם ערכו גבוה מחלק Θ ונמוך מחלק $(1-\Theta)$ של שאר המשתנים, כאשר פונקציה ההתפלגות שלו היא:

$$F(y) = \text{Prob}(Y \leq y)$$

אחוזון ה- Θ של משתנה Y כלשהו מוגדרת כפונקציה ההפוכה, כאשר $0 < \Theta < 1$:

$$Q(\Theta) = \inf(y: F(y) \geq \Theta)$$

לחציון, $Q(\Theta) = 1/2$, ישנו תפקיד מרכזי.

נדרשת נקודת ייחוס עבור משתנה רנדומלי

ידוע כי לכל דגימה $\{y_1, \dots, y_n\}$ של Y , החציון הוא מינימום הערך המוחלט של הסטיות:

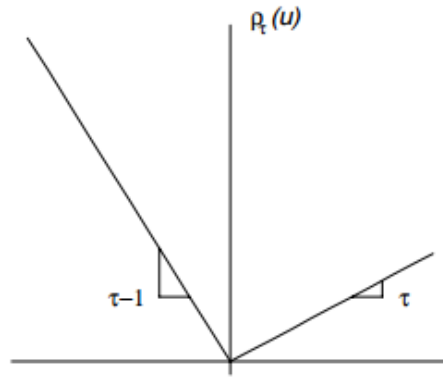
$$\min_{\varepsilon \in R} \sum_{i=1}^n |y_i - \varepsilon|$$

נציג את פונקציית ההפסד ρ_θ . גם ברגרסיה האחוזונית נרצה להביא את סכום השגיאות למינימום:

$$\rho_\theta(u) = z(\theta - I(u < 0))$$

כאשר I הינה פונקציית אינדיקטור⁴.

⁴ פונקציית אינדיקטור – פונקציה המוגדרת בקבוצה X ומציינת שייכות לתת קבוצה A של X . הפונקציה תהיה שווה ל-1 אם ישנה שייכות ול-0 אם אין שייכות.



שרטוט 1. פונקצית ההפסד - רגרסיה האחוזנית

נרצה למצוא $\hat{x}$ כזה שימזער את פונקצית ההפסד על ידי הבאה למינימום את הביטוי הבא :

$$E[\rho(X - \hat{x})] = (\theta - 1) \int_{-\infty}^{\hat{x}} (X - \hat{x}) dF(x) + \int_{\hat{x}}^{\infty} (X - \hat{x}) dF(x)$$

נגזור ונשווה ל-0 :

$$0 = (1 - \theta) \int_{-\infty}^{\hat{x}} dF(x) - \theta \int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \theta$$

אחוזון ה- θ הכללי $\varepsilon(\theta)$ אשר אנלוגי ל- $Q(\theta)$ יכול להיות מוצג כבעיית האופטימיזציה :

$$\min_{\varepsilon \in R} \sum_{i=1}^n \rho_\theta(y_i - \varepsilon)$$

את האמד לתוחלת, בו משתמשים בשיטת הריבועים המזעריים, אשר ממזער את סכום הפרשי הריבועים :

$$\hat{\mu} = \operatorname{argmin}_{\mu \in R} \sum_{i=1}^n (y_i - \mu)^2$$

ניתן להרחיב לפונקצית התוחלת המותנה : $E(Y|X = x) = x'\beta$ על ידי פתרון :

$$\hat{\beta} = \operatorname{argmin}_{\beta \in R} \sum_{i=1}^n (y_i - x_i'\beta)^2$$

לכל אחוזון θ .

מכאן, חישוב האומדן הליניארי של האחוזון המותנה מחושב בעזרת מזעור הפונקציה :

$$\widehat{\beta}_\theta = \operatorname{argmin}_{\beta \in R} \left[\sum_{i: y_i \geq x_i'\beta} \theta |y_i - x_i'\beta| + \sum_{i: y_i < x_i'\beta} (1 - \theta) |y_i - x_i'\beta| \right]$$

ביטוי זה ניתן לחשב בשיטות אופטימיזציה שונות (Roger Koenker, 2000).

כפי שראינו, שיטות האמידה הליניאריות הקלסיות נותנות את האפשרות לאמוד את התוחלת המותנית של הפונקציה בזמן שהגרסיה האחוזונית מספקת שיטות לאמידת החציון המותנה של הפונקציה.

בעוד הגרסיה הליניארית עונה על השאלה: כיצד התוחלת המותנה של המשתנה התלוי מושפעת מהמשתנים הבלתי תלויים, אומדן הגרסיה האחוזונית עונה על אותה השאלה עבור כל רמה של אחוזון.

4. בעיית האינדוגניות ושיטת IV לפתרון הבעיה:

משתנה מקרי אינדוגני הינו משתנה מסביר מקרי אשר מתואם עם ההפרעה האקראית. משתנה אינדוגני פוגע בעקיבות⁵ האומדים ובנכונות המודל. בכדי לבדוק האם אומדי הריבועים הפחותים עקיבים יש לבדוק האם גודל ההטיה תקטן ותשאף ל-0 כאשר מספר התצפיות שואפות לאינסוף והאם השונות שואפת ל-0.

נראה כי אם ישנו קשר סטטיסטי בין המשתנים המסבירים לבין ההפרעות האקראיות, אזי האמידה בשיטת ה-OLS לא מניבה אומדים עקיבים:

$$\hat{\beta} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'\varepsilon$$

$$plim(\hat{\beta} - \beta) = plim\left(\frac{X'X}{n}\right)^{-1} plim\left(\frac{X'\varepsilon}{n}\right) =$$

$$\Sigma_{XX} \underbrace{Cov(X, \varepsilon)}_{\neq 0}$$

הסיבות העיקריות ליצירת בעיית האינדוגניות יכולות להיות: טעויות במדידת המשתנים במשוואה, משוואות סימולטניות(משוואות התלויות זו בזו והמסבירות זו את זו), משתנה תלוי בפיגור כמשתנה מסביר במשוואה.

נשתמש בשיטת IV על מנת להתמודד עם בעיית האינדוגניות:
השיטה הנפוצה לטיפול בבעיית האינדוגניות היא הכנסת עמודה עם משתני עזר במטריצת המשתנים המסבירים במקום העמודה בה נמצא המשתנה האינדוגני. חייב להיות מתאם חזק בין משתני העזר לבין המשתנים המסבירים המקוריים והיעדר מתאם בין משתני העזר להפרעה האקראית.
במקרה בו ישנו מתאם בין המשתנה המסביר לבין ההפרעה האקראית, האומד בהכרח יהיה מוטה:

$$y_i = \alpha + \beta x_i + u_i$$

$$\text{cov}(x_i, u_i) \neq 0$$

$$\hat{\beta} = \beta + \sum \left(\frac{(x_i - \bar{x})u_i}{\sum (x_i - \bar{x})^2} \right)$$

ישנו מתאם במונה ולכן נשתמש במשנה עזר Z .

⁵ עקיבות – הסתברות אומד עקיב לסטות מהפרמטר האמיתי שואפת ל-0 כאשר מספר התצפיות שואף לאינסוף.

נקבל כי האומד החדש הוא :

$$\hat{\beta}_{IV} = \frac{\sum(z_i - \bar{z})y_i}{\sum(z_i - \bar{z})(x_i - \bar{x})} = \frac{\sum(z_i - \bar{z})(\alpha + \beta x_i + u_i)}{\sum(z_i - \bar{z})(x_i - \bar{x})} = \beta + \frac{\sum(z_i - \bar{z})u_i}{\sum(z_i - \bar{z})(x_i - \bar{x})}$$

ניתן לראות כי תוחלת האומד אינה שווה ל-0 מכיוון שישנו X במכנה. לכן נאמר כי האומד מוטה אך עקיב. מה שאומר שככל שנגדיל את מספר התצפיות נגיע לאומדים מדויקים יותר.

יש לציין כי בחירת משתנה העוזר משנה דרמטית את התוצאות בגלל המתאם השונה הקיים בין X ל- Z .

סימולציות:

מכיוון שברצוננו לבדוק את יעילות הרגרסיה האחוזונית לעומת הרגרסיה הרגילה בניבוי הריבית האופטימלית שעל הבנק לתת ללקוח הלווה ממנו, בשלב הראשון, נבנה ב-MATLAB מדגם של 10,000 חברות הלוויות כסף מהבנק. עבור כל חברה אנו מגרילים שלושה משתנים מסבירים הידועים לבנק:

(1 Firm Age – וותק החברה בשוק

(2 Sales – רווחי החברה

(3 Debts – חובות החברה.

בנוסף, לכל חברה יש את החלק ה-Private שלה (המשתנה החבוי) – טיפוס החברה המוגדר מ-0 עד 1 – המשתנה אותו הבנק לא יודע ואותו נרצה לאמוד ואשר בנוי באופן הבא:

(1 הוא מתואם עם שלושת השיפועים שהם ההשפעות השוליות של המשתנים המסבירים ברגרסיה.

(2 הוא אינו מתואם עם המשתנים המסבירים עצמם.

(3 השיפועים מתואמים עם המשתנים המסיבירים (כל שיפוע מתואם עם המשתנה המסביר שלו).

(4 הוא מתפלג ביטא שהיא התפלגות המתאימה למשתנה שמקבל ערכים בין 0 ל-1.

נעשה שימוש בהתפלגות רב-נורמלית¹, המקיימת את 4 הדרישות לעיל. תחת ההנחה שהחלק היחיד שנמצא בהפרעה האקראית ומתואם עם השיפועים ברגרסיה הוא האינפורמציה הפרטית.

השיפועים של משתנה ה-Private הינם משתנים מקריים בעצמם (לכל אחד מהם הנחנו התפלגות שולית נורמלית כיוון שההתפלגות של כל הנתונים היא רב-נורמלית) והם מתואמים באופן שיטתי עם מרכיב שנמצא בהפרעה האקראית (החלק הבלתי נצפה של Private).

אם היה ידוע לבנק טיפוס החברה, היה יכול לשחזר במדויק את הריבית האופטימלית עליו להציע לחברה, אך משום שהבנק משתמש באינפורמציה הגלומה במשתנים המסבירים הוא יכול לחשב רק חלק מהאינפורמציה על טיפוס החברה ולא יידע לחשב את ההפרעה. לבנק יש רק התפלגות אפריורית של הטיפוסים של הלווים, הוא אינו יכול לצפות בחלק ה-Private של הלווה.

בשלב השני, נבנה את הפונקציות הבאות:

הריבית הבסיסית – אותה החברה משלמת בפועל – הינה פונקציה ליניארית של המשתנים המסבירים הידועים לבנק:

¹ התפלגות רב-נורמלית - הכללה של התפלגות נורמלית למשתנים מקריים רב-ממדיים.

הריבית האופטימלית – הריבית אותה הייתה משלמת החברה אם היה ידוע לבנק החלק ה-Private שלה – הינה פונקציה ליניארית של הריבית הבסיסית יחד עם המשתנה החבוי ושגיאה אקראית.

פונקציית ההפסד של הבנק, אותה היא רוצה למזער, תהיה הפער (בריבוע) בין הריבית שהיה ראוי לתת לפירמה אם הטיפוס שלה היה ידוע, לבין הריבית שהפירמה משלמת בפועל.

הרווח של הבנק יהיה התוחלת של תקבולי הריבית פחות פונקציית ההפסד.

היה חשוב מאוד לבדוק כי אכן אין אנו נתקלים בבעיית האינדוגניות ושהאומדים הינם עקיבים, לכן בשלב השלישי בדקנו כי אין תלות בין המשוואות, וכי אין מתאם בין המשתנים המסבירים לבין ההפרעה האקראית.

מטרת הסימולציות הינה למזער את פונקציית ההפסד של הבנק פעם בשיטת הרגרסיה הליניארית הרגילה ופעם בשיטת הרגרסיה האחוזונית ולבדוק באיזו שיטה יתקרב הבנק כמה שיותר לריבית האופטימלית.

לכן בשלב הרביעי, נבצע את הסימולציות עבור שתי השיטות שהצגנו :

בשיטת הרגרסיה הליניארית הקלסית :

נאמוד את הרגרסיה ונחשב את התחזיות על הריבית עבור שלושת המשתנים המסבירים הידועים לבנק, כאשר המשתנה התלוי יהיה הריבית האופטימלית, ונחשב את סכום השגיאות הריבועיות.

בשיטת הרגרסיה האחוזונית :

נחלק את המשתנה התלוי(הריבית האופטימלית) לאחוזונים [5,10,15,...95] ונחשב את תחזית הריביות עבור כל אחד מהאחוזונים. לבסוף נחשב את סכום השגיאות הריבועיות.

בשלב האחרון, נבדוק כיצד משתנה הפער בין סכום השגיאות הנמדד בשיטת הרגרסיה הליניארית הרגילה לבין סכום השגיאות הנמדד בשיטת הרגרסיה האחוזונית, כאשר אנו משנים את מקדמי המשתנים המסבירים ואת מקדם המשתנה החבוי של החברות ונפיק את הגרפים המציגים זאת. מקדמי המשתנים משתנים בין 0.01 עד 1 בקפיצות של 0.09.

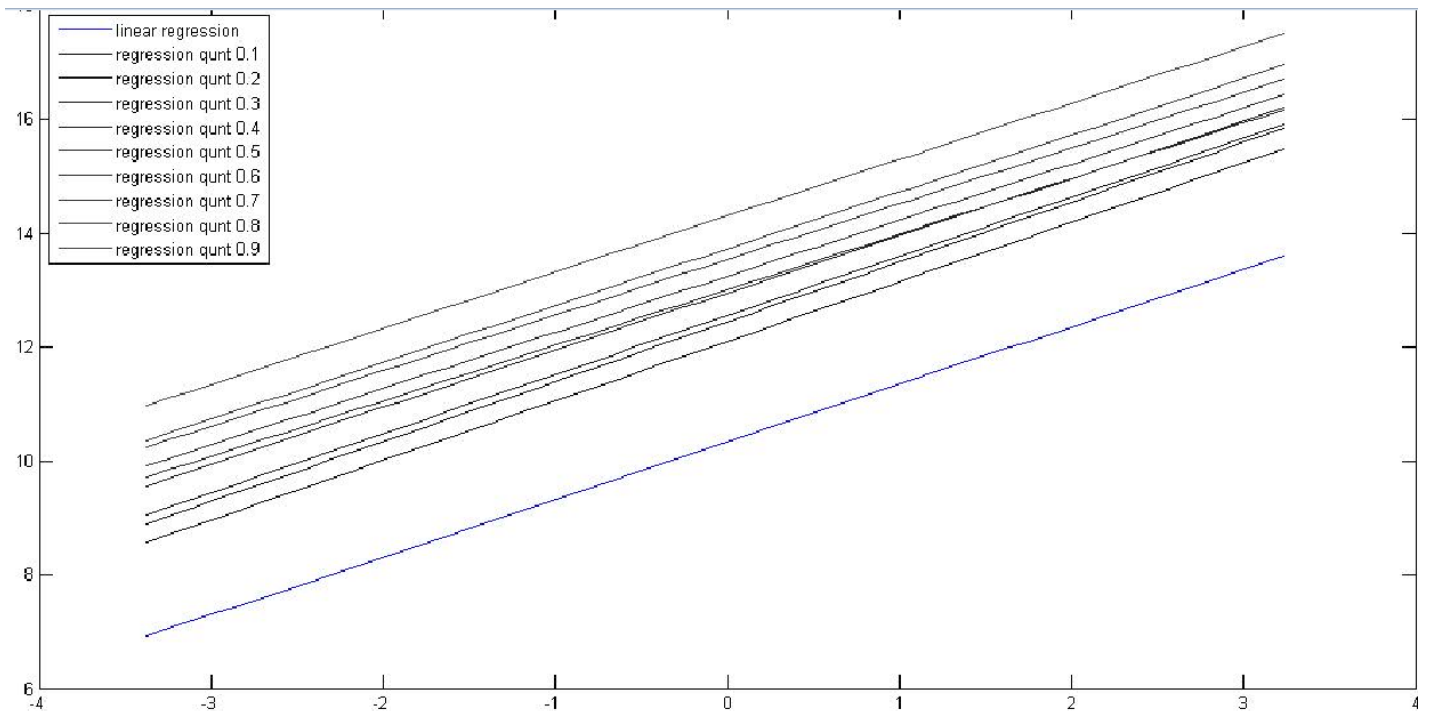
תוצאות הסימולציה ומסקנות:

הטבלה שלהלן מציגה את מקדמי הרגרסיה הרגילה לעומת האחוזונית ומשווה ביניהם:

	regular regression	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Offset	4.645025311	3.876925	4.158946	4.320607	4.472989	4.636135	4.778053	4.936565	5.115181	5.431873	7.115492
Firm Age	0.954164429	0.91311	0.891988	0.897569	0.918015	0.938248	0.946384	0.954828	0.965932	1.010059	0.914616
Distance to default	0.983703923	0.982667	0.988696	0.967506	0.946618	0.953987	0.968127	0.994115	1.027981	1.001307	0.703891
Sales	1.468805572	1.471426	1.471046	1.473378	1.473509	1.474085	1.492913	1.468269	1.448982	1.465929	1.254009

טבלה 1. השפעת המשתנים בכל אחוזון

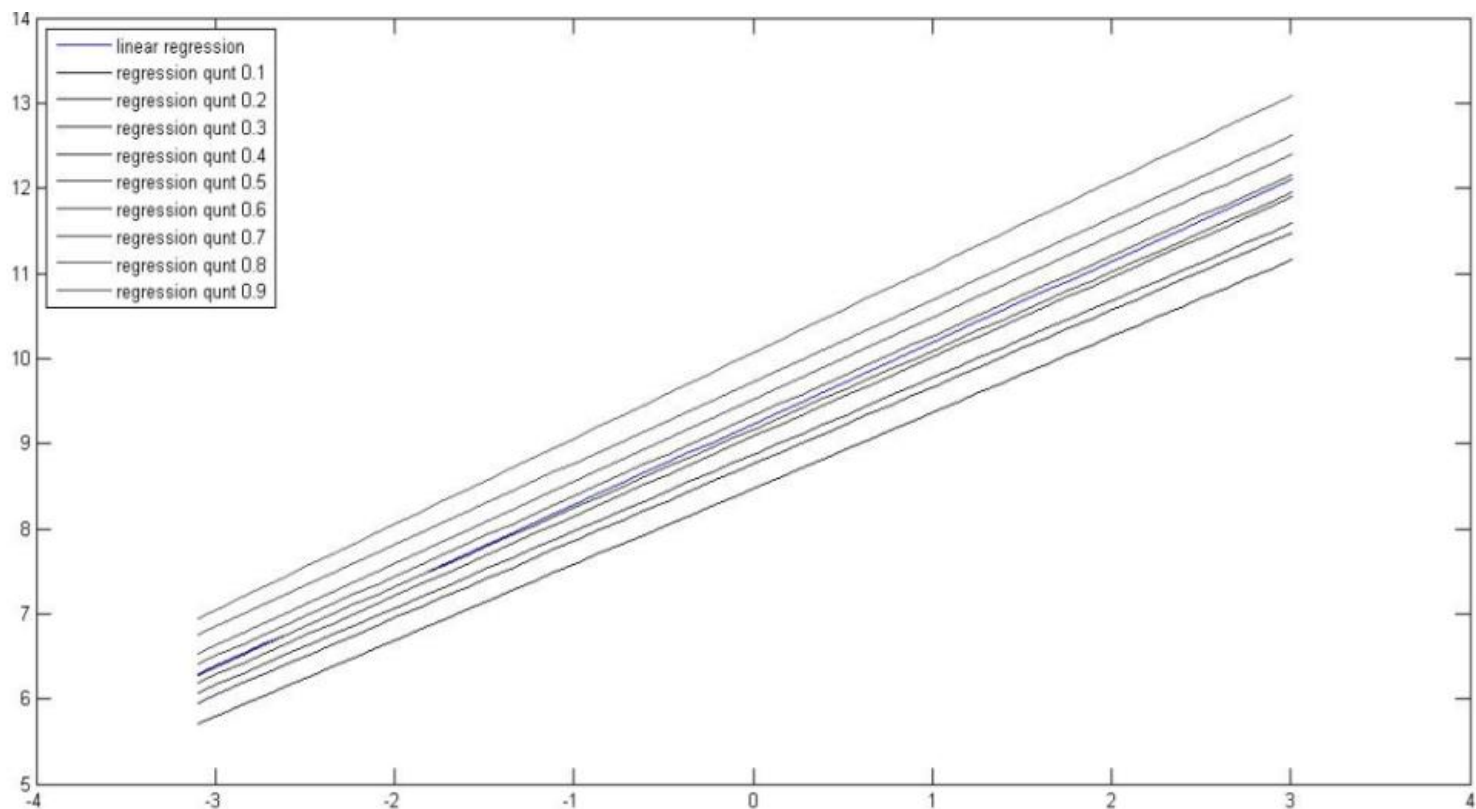
הגרפים הבאים מציגים את קווי הרגרסיה המתקבלים בשיטה הליניארית הרגילה ובשיטה האחוזונית עבור משתנה מסביר אחר כל פעם, כאשר הקו הכחול הינו קו הרגרסיה הליניארית הרגילה:



גרף 1. קווי הרגרסיה במשתנה Debts

Firm age=2

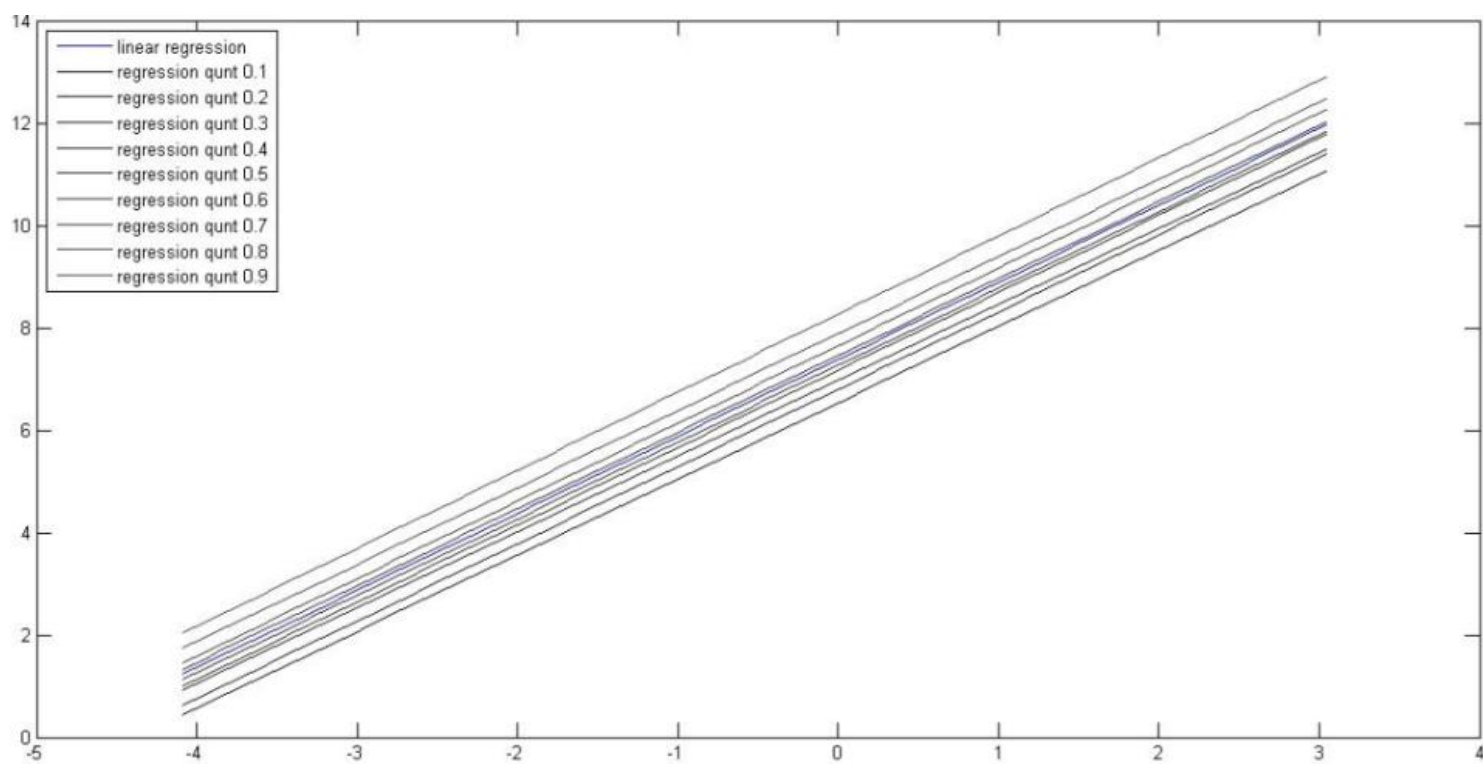
Sales=2.5



גרף 2. קווי הרגרסיה במשתנה Firm age

Debts =0.8

Sales=2.5



גרף 3. קווי הרגרסיה במשתנה Sales

Debts =0.8

Firm age=2

- כאשר בדקנו את סכום הריבועים המזעריים בשתי השיטות – שהוא בעצם הפער בין האומדים לבין התוצאות בפועל – קיבלנו כי שיטת הרגרסיה הליניארית הקלסית סכום השגיאות הוא : 393.2 לעומת הרגרסיה האחוזונית בה סכום השגיאות הינו 52.3 בסה"כ.

מסקנות :

- על פי טבלה 1 ניתן לראות כי משקל המשתנה המסביר משתנה לפי אחוזונים. לדוגמא ניתן לראות כי באחוזון ה-100 גודל החובות שיש לחברה מסויימת משפיע הרבה פחות מבאחוזונים הנמוכים יותר על הריבית האופטימלית שיש לתת לה. אכן אפשר לראות כי לאורך האחוזונים ישנו שינוי בהשפעת המשתנים על קווי הרגרסיה.
- גרפים 1-3 מראים בצורה מובהקת כיצד משתנה קו הרגרסיה בכל אחוזון וכיצד קווי הרגרסיה שונים מקו הרגרסיה המופק בשיטת הרגרסיה הליניארית הקלסית.
- ניתן לומר כי התוצאות שונות מאוד בכל אחוזון וההבדלים המהותיים שיש בין הריבית האופטימלית המוצעת על ידי שיטת הרגרסיה הליניארית הקלסית לבין האחוזונית יכולים לפגוע קשות בהסקת המסקנות לגבי הריבית האופטימלית המגיעה ללקוח. בגרף 1 קו הרגרסיה המופק בשיטה הרגילה נמצא מובדל לחלוטין מקווי הרגרסיה האחוזונית והשימוש ברגרסיה המוצעת בדרך זו גוררת טעות קריטית בהצעת ריבית אופטימלית ללקוח, הריבית האופטימלית שתוצע תהיה נמוכה מדי אפילו לגבי האחוזון הנמוך ביותר של החברות, מה שיגרום להפסדים קשים לבנק.
- בגרפים 2-3 התקבלו תוצאות צמודות יותר, כאשר קו הרגרסיה המופק על ידי הרגרסיה הרגילה כמעט מתלכד עם האחוזון ה-50%. ולמרות זאת, גם כאן ניתן לומר כי קביעת הריבית האופטימלית תהיה מדויקת ונכונה יותר אם נלך על פי שיטת הרגרסיה האחוזונית.
- סכום הריבועים המזעריים קטן בצורה משמעותית כאשר מדובר ברגרסיה האחוזונית ומראה כי אכן השימוש בשיטה זו מניב אומדים המתקרבים יותר לערכים האמיתיים.

סיכום ומחקרי המשך:

בעבודה זו נעשה ניסיון לבדוק איזו שיטה יעילה יותר לאומדן ריבית אופטימלית כאשר המשתנים המסבירים מרובים: שיטת הרגרסיה הליניארית הקלסית, לעומת שיטת הרגרסיה האחוזונית.

כאמור, ממחקרים קודמים עולה כי שיטת הרגרסיה האחוזונית מצליחה להסביר טוב יותר את הקשרים בין המשתנים ועמידה יותר לערכים קיצוניים. כמו כן, החלוקה לאחוזונים מאפשרת ראייה רחבה יותר על השפעת המשתנים המסבירים בכל רמת אחוזון וניתן להסיק מסקנות ספציפיות על כל אחד מהם, מה שתורם גם ליכולת הניבוי של החוקר.

לכן ההנחה הייתה כי נוכל להתקרב טוב יותר לערכים האמיתיים של הריבית האופטימלית דווקא על ידי שימוש בשיטה האחוזונית. לשם כך בנינו מדגם של 10,000 חברות ומשוואת רגרסיה אותה אמדנו פעמיים: פעם אחת על ידי השיטה הליניארית הרגילה ופעם אחרת על ידי השיטה האחוזונית.

השימוש בשיטה האחוזונית עוקף את הצורך בשימוש במשתני דמי. משתני דמי אשר נבחרים בהתאם לערך המשתנה התלוי ואשר מטרתם לייצר שיפוע שונה לכל תת קבוצה של תצפיות וכן חותך שונה כאילו הרצנו רגרסיה נפרדת לכל תת מדגם, תיצור בעיית אינדוגניות מכיוון שבצורה זו נוצרת סיבתיות הפוכה. השימוש ברגרסיה אחוזונית מצליחה להתגבר על הבעיה מכיוון שהחלק ה-Private אותו אנו אומדים נוצר כך שהוא בנוי ממשתנים מקריים אשר אינם מתואמים עם ההפרעה האקראית של הרגרסיה עצמה אך מתואם עם המשתנים המסבירים האחרים.

בדקנו עבור מקרה ספציפי כלשהו מה הם האומדים למקדמי הרגרסיה המתקבלים בשתי השיטות. קיבלנו כי הרגרסיה האחוזונית מציגה שינוי באומדים לאורך כל האחוזונים בכל משתנה מסביר לעומת הרגרסיה הליניארית הרגילה אשר מציגה אומדים המכלילים בתוכם את התוצאות לאורך האחוזונים מה שמסביר בצורה כללית יותר את התוצאות וקשה לראות את המגמות כפי שניתן לעשות עבור כל אחוזון.

יצרנו גרפים המציגים את קווי הרגרסיה המתקבלים בשיטה הליניארית הרגילה ובשיטה האחוזונית עבור משתנה מסביר אחר כל פעם וראינו כי ישנם הבדלים מהותיים בין קווי הרגרסיה באחוזונים השונים לבין קו הרגרסיה המופק בשיטת הרגרסיה הליניארית הקלסית.

ניתן לומר, כי הצלחנו להתמודד עם שיפועים המשתנים על פני קבוצות תצפיות, זאת מבלי להכניס למודל אינדוגניות, דבר שרגרסיה רגילה לא היתה מצליחה ליישם. ולמעשה הצלחנו להקטין את השגיאה בחיזוי המשתנה התלוי, זאת למרות שאנו לא צופים בחלק ה-Private.

מסקנתנו הייתה כי האמידה בשיטת הרגרסיה האחוזונית אכן משפרת דרסטית את הפער בין האומדים לבין התוצאות בפועל מה שמאפשר ניבוי מדויק יותר בעתיד ומציג תמונה רחבה וברורה יותר על השפעת המשתנים השונים על התוצאות.

מחקרי המשך:

- המדגם אשר נלקח היה אקראי ויהיה זה מעניין לבדוק את התוצאות עבור מדגם ריאלי הנלקח מנתונים אמיתיים של חברות קיימות.
- במקרה שלנו בדקנו את השפעת החלוקה לאחוזונים במקרה הליניארי, במחקרי ההמשך ניתן לבדוק כיצד משפיעה החלוקה על מקרים שאינם ליניאריים.

רשימה בבליוגרפית:

- 1) Weerts, A.H. · Winsemius, H.C. · Verkade, J.S. (2011). Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales)
- 2) Simona Despa. (2007). Stat news #70: Quantile regression – Cornell University.
- 3) Roger Koenker. (2000). "Quantile regression" - Department of Economics University of Illinois, Urbana-Champaign Champaign, IL 61820 USA.
- 4) Moshe Buchinsky. (1994). Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression
- 5) Xuming He, Ying Wei. (2005). Analysis of Panel Data. Second edition. Chang Hsiao, University of southern California.
- 6) Roger Koenker and Kevin F. Hallock. (2001). Quantile Regression. Journal of Economic Perspectives—Volume 15. Pages 143–156.
- 7) Marco Avellaneda and Jingyi Zhu. (2001). Distance to default.
- 8) גדעון מאור שביט ומירי סבג-אנדבלד. (2004). מדידת פערי שכר בין המגזר הציבורי למגזר הפרטי: ממצאים משימוש ברגרסיה אחוזהנית.
- 9) ד"ר דרורה קרוטקין. אקונומטריקה – חומר תאורטי.

```

MU=[0 0 0 0]; rho=0.4;
Sigma=ReshapeSig([1 rho 0 0 0 1 0.6 0.2 0.1 1 0.2 -0.3 1 0.3 1], 5 );
n=1000;
W=mvnrnd(MU,Sigma,n);
Type = betainv(normcdf(W(:,1)),2,2); B1=2+0.5*W(:,2); X1=W(:,3); X23=W(:,4:5);
eps=normrnd(0,1,n,1);
% interest rate:
Y=4+0.5*(B1.*X1+ X23*[2;3]+2*Type+eps);

% percentile of slope 1 (b1):
prctile(B1,[10 20 30 40 50 60 70 80 90])
corr(eps,[X1 X23])
corr(Type,X1)
corr(Type,B1)

%BANK ASSUMPTION
data_for_regress = [ones(size(X1)) X1 X23];
[regress_coefs, regress_coefs_confidance, r_regress] = regress(Y,data_for_regress);
regular_coef_vec = regress_coefs;
%devide the data into 20 quantiles by Ropt and return the quantile boarders
quantile_Ropt_vec = quantile(Y,9)';
%quantile_mat columns are indices of firms of the relevant quantile
%(according to quantile_Ropt_vec)
number_of_quantiles = numel(quantile_Ropt_vec)+1;

firms_per_quantile = n/number_of_quantiles;

quantile_mat = zeros(firms_per_quantile,number_of_quantiles);
% sort the different borrowers to their quantiles by index
A=find(Y<quantile_Ropt_vec(1));
quantile_mat(:,1) = A; %[A;zeros(firm_num-numel(A),1)]
for i=2:numel(quantile_Ropt_vec)
    A=find(Y>quantile_Ropt_vec(i-1) & Y<quantile_Ropt_vec(i));
    quantile_mat(:,i) = A;
end
A=find(Y>quantile_Ropt_vec(end));
quantile_mat(:,end) = A;

% quantile regression - regress for every quantile individually

data_for_quantreg = zeros(firms_per_quantile,4,number_of_quantiles);
quantreg_coefs = zeros(4,number_of_quantiles);
quantreg_coefs_confidance = zeros(4,2,number_of_quantiles);
r_quantreg = zeros(firms_per_quantile,number_of_quantiles);

figure(1);
Sales_vec = linspace(min(X1),max(X1),100);
FirmAge_vec = linspace(min(X23(:,1)),max(X23(:,1)),50)
Dist2def_vec = linspace(min(X23(:,2)),max(X23(:,2)),50);

```

```

regressSales = regress_coefs(1) + regress_coefs(2)*2 + regress_coefs(3)*0.8 +
regress_coefs(4)*Sales_vec;
regressFA = regress_coefs(1) + regress_coefs(2)*FirmAge_vec + regress_coefs(3)*0.8 +
regress_coefs(4)*2.5;
regressDist = regress_coefs(1) + regress_coefs(2)*2 + regress_coefs(3)*Dist2def_vec +
regress_coefs(4)*2.5;
plot (Sales_vec,regressSales,'b'); hold on;
title(['Regression Sales' char(10) 'Distance to Default = 0.8' char(10) 'Firm Age =
2']);
figure(2);
plot (FirmAge_vec,regressFA,'b'); hold on;
title(['Regression Firm Age' char(10) 'Distance to Default = 0.8' char(10) 'Sales =
2.5']);
figure(3);
plot (Dist2def_vec,regressDist,'b'); hold on;
title(['Regression Distance to Default' char(10) 'Firm Age = 2' char(10) 'Sales =
2.5']);
for i= 1 : number_of_quantiles - 1
    data_for_quantreg(:, :, i) = [ones(firms_per_quantile,1) x1(quantile_mat(:, i))
x23(quantile_mat(:, i), :)];
    [quantreg_coefs(:, i), quantreg_coefs_confidance(:, :, i), r_quantreg(:, i)] =
regress(Y(quantile_mat(:, i)), data_for_quantreg(:, :, i));

    [p, stats]=quantreg(data_for_regress, Y, 0.1*i)
    quantreg_coefs(:, i) = p;
    %return

    regressSales = quantreg_coefs(1, i) + quantreg_coefs(2, i)*2 +
quantreg_coefs(3, i)*0.8 + quantreg_coefs(4, i)*Sales_vec;
    regressFA = quantreg_coefs(1, i) + quantreg_coefs(2, i)*FirmAge_vec +
quantreg_coefs(3, i)*0.8 + quantreg_coefs(4, i)*2.5;
    regressDist = quantreg_coefs(1, i) + quantreg_coefs(2, i)*5 +
quantreg_coefs(3, i)*Dist2def_vec + quantreg_coefs(4, i)*2.5;
    figure(1); plot (Sales_vec, regressSales, 'color', [0.03*i 0.03*i 0.03*i]); hold on;
    figure(2); plot (FirmAge_vec, regressFA, 'color', [0.03*i 0.03*i 0.03*i]); hold on;
    figure(3); plot (Dist2def_vec, regressDist, 'color', [0.03*i 0.03*i 0.03*i]); hold
on;
end
[p, stats]=quantreg(data_for_regress, Y, 0.999)
quantreg_coefs(:, end) = p;
quantreg_coef_mat = quantreg_coefs;

Legend=cell(number_of_quantiles+1, 1);
Legend{1}='linear regression';
for iter=2:number_of_quantiles+1
    Legend{iter}=['regression quant ' num2str(10*(iter-1)/100)];
end
figure(1); legend(Legend, 'Location', 'northwest');
figure(2); legend(Legend, 'Location', 'northwest');
figure(3); legend(Legend, 'Location', 'northwest');

Rsquared_regress = sum(r_regress.^2);
Rsquared_quantreg = sum(r_quantreg(:).^2);

r_sqr_diff = (Rsquared_regress - Rsquared_quantreg)/Rsquared_regress;

```

```

MU=[0 0 0 0 0]; rho=0.4;
Sigma=ReshapeSig([1 rho 0 0 0 1 0.6 0.2 0.1 1 0.2 -0.3 1 0.3 1], 5 );
n=1000;
w=mvnrnd(MU,Sigma,n);
Type = betainv(normcdf(w(:,1)),2,2); B1=2+0.5*w(:,2); X1=w(:,3); X23=w(:,4:5);
eps=normrnd(0,1,n,1);
% interest rate:
Y=4+0.5*(B1.*X1+ X23*[2;3]+2*Type+eps);

% percentile of slope 1 (b1):
prctile(B1,[10 20 30 40 50 60 70 80 90])
corr(eps,[X1 X23])
corr(Type,X1)
corr(Type,B1)

%BANK ASSUMPTION
data_for_regress = [ones(size(X1)) X1 X23];
[regress_coefs, regress_coefs_confidance, r_regress] = regress(Y,data_for_regress);
regular_coef_vec = regress_coefs;
%divide the data into 20 quantiles by Ropt and return the quantile boarders
quantile_Ropt_vec = quantile(Y,9)';
%quantile_mat columns are indices of firms of the relevant quantile
%(according to quantile_Ropt_vec)
number_of_quantiles = numel(quantile_Ropt_vec)+1;

firms_per_quantile = n/number_of_quantiles;

quantile_mat = zeros(firms_per_quantile,number_of_quantiles);
% sort the different borrowers to their quantiles by index
A=find(Y<quantile_Ropt_vec(1));
quantile_mat(:,1) = A; %[A;zeros(firm_num-numel(A),1)]
for i=2:numel(quantile_Ropt_vec)
    A=find(Y>quantile_Ropt_vec(i-1) & Y<quantile_Ropt_vec(i));
    quantile_mat(:,i) = A;
end
A=find(Y>quantile_Ropt_vec(end));
quantile_mat(:,end) = A;

% quantile regression - regress for every quantile individually

data_for_quantreg = zeros(firms_per_quantile,4,number_of_quantiles);
quantreg_coefs = zeros(4,number_of_quantiles);
quantreg_coefs_confidance = zeros(4,2,number_of_quantiles);
r_quantreg = zeros(firms_per_quantile,number_of_quantiles);

figure(1);
Sales_vec = linspace(min(X1),max(X1),100);
FirmAge_vec = linspace(min(X23(:,1)),max(X23(:,1)),50)
Dist2def_vec = linspace(min(X23(:,2)),max(X23(:,2)),50);
regressSales = regress_coefs(1) + regress_coefs(2)*2 + regress_coefs(3)*0.8 +
regress_coefs(4)*Sales_vec;
regressFA = regress_coefs(1) + regress_coefs(2)*FirmAge_vec + regress_coefs(3)*0.8 +
regress_coefs(4)*2.5;
regressDist = regress_coefs(1) + regress_coefs(2)*2 + regress_coefs(3)*Dist2def_vec +
regress_coefs(4)*2.5;

```

```

plot (Sales_vec,regressSales,'b'); hold on;
title(['Regression Sales' char(10) 'Distance to Default = 0.8' char(10) 'Firm Age = 2']);
figure(2);
plot (FirmAge_vec,regressFA,'b'); hold on;
title(['Regression Firm Age' char(10) 'Distance to Default = 0.8' char(10) 'Sales = 2.5']);
figure(3);
plot (Dist2def_vec,regressDist,'b'); hold on;
title(['Regression Distance to Default' char(10) 'Firm Age = 2' char(10) 'Sales = 2.5']);
for i= 1 : number_of_quantiles - 1
    data_for_quantreg(:,i) = [ones(firms_per_quantile,1) x1(quantile_mat(:,i))
x23(quantile_mat(:,i),:)]';
    [quantreg_coefs(:,i), quantreg_coefs_confidance(:,i), r_quantreg(:,i)] =
regress(Y(quantile_mat(:,i)),data_for_quantreg(:,i));

    [p,stats]=quantreg(data_for_regress,Y,0.1*i)
    quantreg_coefs(:,i) = p;
    %return

    regressSales = quantreg_coefs(1,i) + quantreg_coefs(2,i)*2 +
quantreg_coefs(3,i)*0.8 + quantreg_coefs(4,i)*Sales_vec;
    regressFA = quantreg_coefs(1,i) + quantreg_coefs(2,i)*FirmAge_vec +
quantreg_coefs(3,i)*0.8 + quantreg_coefs(4,i)*2.5;
    regressDist = quantreg_coefs(1,i) + quantreg_coefs(2,i)*5 +
quantreg_coefs(3,i)*Dist2def_vec + quantreg_coefs(4,i)*2.5;
    figure(1); plot (Sales_vec,regressSales,'Color',[0.03*i 0.03*i 0.03*i]); hold on;
    figure(2); plot (FirmAge_vec,regressFA,'Color',[0.03*i 0.03*i 0.03*i]); hold on;
    figure(3); plot (Dist2def_vec,regressDist,'Color',[0.03*i 0.03*i 0.03*i]); hold
on;
end
[p,stats]=quantreg(data_for_regress,Y,0.999)
quantreg_coefs(:,end) = p;
quantreg_coef_mat = quantreg_coefs;

Legend=cell(number_of_quantiles+1,1);
Legend{1}='linear regression';
for iter=2:number_of_quantiles+1
    Legend{iter}=['regression qunt ' num2str(10*(iter-1)/100)];
end
figure(1); legend(Legend,'Location','northwest');
figure(2); legend(Legend,'Location','northwest');
figure(3); legend(Legend,'Location','northwest');

Rsquared_regress = sum(r_regress.^2);
Rsquared_quantreg = sum(r_quantreg(:).^2);

r_sqr_diff = (Rsquared_regress - Rsquared_quantreg)/Rsquared_regress;

```

```
function S = ReshapeSig(t,size_)
h = 1;
S = zeros(size_, size_);
for i=1:size_
    for j=i:size_
        S(i,j) = t(h);
        S(j,i) = t(h);
        h = h + 1;
    end
    % [t(1) t(2) t(3); t(2) t(4) t(5); t(3) t(5) t(6)]
end
end
```