|  |  |
| --- | --- |
| המחלקה לכלכלה  אוניברסיטת בן-גוריון בנגב  מבוא לאקונומטריקה 142-1-1081  שנה"ל תשפ"א | The Economics Department  Ben-Gurion University of the Negev  Introductory Econometrics 142-1-1081  Academic year 2020-1 |
| **פרויקט במבוא לאקונומטריקה** | |

חלק א: תיאור המודל והנתונים.

1. **Hypothesis:**

בעבודה שלנו, אנחנו מנסים להבין ולהסביר את הדרוג של אנימה ומהם הגורמים שיכולים להשפיע עליו. נבדוק את הקשר בין הגורמים שונים כמו סוג הצפייה, אורך זמן, כמות האזהרות וכו‘ לבין ההערכה והדרוג לאנימה. הבנה של ההשפעה של המשתנים השונים על האנימה יכולה לעזור למשקיעים ובעלי העסק להשקיע יותר כסף וזמן באנימה שממקסם כמה שיותר רווחים והנאה של קהל היעד המעוניין בצפייה באנימה.

1. **Data sources:**

שם בסיס הנתונים שננתח אותו הוא anime. את בסיס הנתונים הזה נלקח מאתר [Kaggle](https://www.kaggle.com/alancmathew/anime-dataset), ונוצר על ידי Alan Mathew, סטודנט באוניברסיטת ווטרלו בקנדה. הוא אסף את הנתונים מהאתר [anime-planet](https://www.anime-planet.com/) בתאריך 5-6-2020, נתונים שקשורים לכל אנימה, כמו השם, תיאור, מספר צפיות, מספר מצביעים וכו'.

1. **Dataset:**

תארו את מבנה הנתונים שלכם: נתוני חתך (cross-section) / נתונים על פני זמן (time-series) וציינו מה מייצגת תצפית *i/t*.

מבנה הנתונים מורכב מנתוני חתך. בסיס הנתונים הנבחר לא נכלל בתוכו נתונים על פני זמן.

כל תצפית i בסט הנתונים מתארת אנימה מסוים, תכונות ונתונים שקשורים ספציפית אליו.

לדוגמה, עבור התצפית השנייה(i=2\אנימה שני) בסט הנתונים:



1. - משתנים רציפים:
   * משתנה 1, **streaming\_time**: משתנה זה מייצג את מספר השנים שבו האנימה הוצג בהם. ). לדוגמה, האנימה A Silent Voice, הופץ בשנת 2016, והסתיים באותה שנה. כלומר streaming\_time = 0.
   * משתנה 2, **tags\_count**: במשתנה הזה, אנחנו סופרים את כמות התגים עבור כל אנימה(תצפית). לדוגמה, עבור האנימה A Silent Voice, יש לו את התגים האלה: ['Drama', 'Shounen', 'Disability', 'Melancholy', 'Mental Illness', 'School Life', 'Based on a Manga'], ולכן המשתנה tags\_count יהיה שווה ל-7.
   * משתנה 3, **count\_warning**: המשתנה הנ"ל, סוכמים את כמות האזהרות שמופיעות עבור כל אנימה. דוגמה: לאנימה A Silent Voice, יש את האזהרות האלה: ['Bullying', 'Mature Themes', 'Suicide'], ולכן tags\_count = 3.

* משתנים קטגוריאליים:
  + משתנה 4, **is\_description**: המשתנה הזה הינו משתנה בינארי(כלומר, משתנה דמה) שמקבל 1 אם לאנימה יש תיאור, ו-0 אין תיאור. לדוגמה, עבור האנימה A Silent Voice, is\_description =1 כי יש לו את התיאור העצוב: " After transferring into a new school, a deaf girl, Shouko Nishimiya, is bullied by the popular Shouya Ishida. As Shouya continues to bully Shoko, the class turns its back on him. Shouko transfers and Shoya grows up as an outcast. Alone and depressed, the regretful Shouya finds Shouko to make amends.Â ".

1. **The Econometric Model**:

המודל האקונומטרי שבחרנו לאמוד:

הערה חשובה: בחלק הבא אנחנו נציג קשר בין משתנה נגד המשתנה , בחרנו להוריד את המשתנה מסיבות שנציג אותם בחלק ב’.

1. **תיאור המשתנים במודל**:

|  |  |  |  |
| --- | --- | --- | --- |
|  | שם המשתנה | מייצג ומודד את: | יחידות המדידה |
| משתנה  מוסבר | rating | הערכת האנימה והדרוג | [0,5] |
| משתני דמה  (קטגוריאליים \ בינאריים) | mediaType | סוג צפייה: הסוגים השונים שבהם ניתן לצפות באנימה. | [Movie, TV Special, Web, OVA, DVD Special, TV, Music Video] |
| is\_description | הסבר לאנימה: האם מצורף לאנימה תיאור או הסבר. | [No,Yes] |
| משתנים רציפים | duration | אורך זמן של אנימה: זמן בדקות | [1,2,…] |
| tags\_count | מספר התגים: מספר התגים המצורפים(action, romance…) | [0,1,2,…) |
| count\_warning | מספר האזהרות המצורפות | [0,1,2,…) |
| votes | מספר המצביעים | [0,1,2,…) |

חלק ב: סטטיסטיקה תיאורית ובניית בסיס הנתונים.

1. **Summary Statistics** (4 נק)

הערה: את המאפיינים הסטטיסטיים ניתן ליישם רק על המשתנים הרציפים במודל.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Table 1. Summary Statistics* | | | | | |
| **max** | **min** | **StDev** | **mean** | **N** | Variables |
| 4.663 | 0.844 | 0.834 | 2.754 | 6830 | rating |
| 163 | 1 | 32.256 | 25.75 | 6830 | duration |
| 20 | 1 | 2.689 | 4.539 | 6830 | tags\_count |
| 7 | 0 | 0.617 | 0.165 | 6830 | count\_warning |
| 82752 | 10 | 3170.3 | 1015 | 6830 | votes |
| 115949 | 0 | 4778.405 | 1572.3 | 6830 | watched |

1. 2 **גרפים** שמתארים התנהגות **של** משתנים במודל:
2. Chart, histogram

   Description automatically generatedהיסטוגרמה של המשתנה votes(משתנה רציף).

ניתן לראות בקלות שההתפלגות של המשתנה הינה עם זנב כבד (heavy-tailed). התאמת רגרסיה לינארית סטנדרטית לנתונים עם זנב כבד מדי תביא לכך שנקודות נתונים בזנבות ייענשו יותר מדי, אך אומדני המקדם במודל עדיין בדרך כלל סבירים. החיסרון העיקרי במצב הזה הוא שמרווחי החיזוי של הערכים יהיו קצרים מדי מכיוון שהם לא לוקחים בחשבון את הזנבות הכבדים. אנחנו לא נעשה תיקון למשתנה(log למשל) מכיוון שזה יגרור שתצפיות יקבלו ערכים ששואפים לאינסוף ונצטרך להוריד את התצפיות האלה.

Chart, treemap chart

Description automatically generated

1. bar plot של המשתנה is\_description ( משתנה בינארי).

היה מוזר לראות שיותר מ- 50% מסט הנתונים הינו על אנימה שאין להם הסבר. הדעה שלנו הייתה שהכנסת הסבר לאנימה יכולה להשפיע חיובית על הדרוג שלו.

1. **2 גרפים** שמתארים קשרים **בין** משתנים במודל:

Chart, scatter chart

Description automatically generated

1. משך הזמן של האנימה נגד הדרוג שלו.

כמו שאנחנו רואים, ככל שהאנימה ארוך יותר, הדרוג שלו יהיה יותר גבוה. הקשר הזה לא היה טריוויאלי, כי ייתכן למשל שככל שהאנימה ארוך יותר, הצופים יהיו משועממים יותר ולכן ייתנו דרוג יותר נמוך, אבל קיבלנו דווקא את הקשר ההפוך, כלומר קשר חיובי.

ניתן לראות גם שמקדם המתאם שווה ל-0.48, ואפשר להגיד על זה שהמשתנה duration יהיה מובהק במודל שנבנה אותו בחלק הבא.

Chart, scatter chart

Description automatically generated

1. מספר המצביעים נגד מספר המצביעים עבור כל אנימה.

אפשר לראות שיש קשר ליניארי חיובי וחזק מאוד בין שני המשתנים, עם מקדם מתאם ששווה ל-0.99. קשר חזק וגבוה מאוד בין שני משתנים אינו דבר טוב, ואפילו דבר שכן נרצה שלא יקרה. הקשר הזה יגרור לנו לבעיה של מולטיקולינאריות, זה גורר לבעיה בחישוב המטריצה כך שלא יהיה לה הופכית ולכן אין פתרון למשוואות הנורמליות. נשאיר רק אחד המשתנים כתוצאה למה שקיבלנו, ונשאיר את המשתנה votes.

חלק ג: אמידת המודל ותוצאות (45 נק)

1. (5 נק) אמדו את המודל ב OLS ודווחו על המקדמים שקיבלתם (estimates), על סטיות התקן (s.e.) שלהם, מקדם ההסבר המתוקן של המודל (R-adjusted) ומספר התצפיות (N) בטבלה שבה המשתנים המסבירים מופיעים בשורות והמשתנים המוסברים של מודלים שונים בעמודות. *דוגמה לטבלה*:

|  |  |  |
| --- | --- | --- |
|  |  | Table 2. Regression results |
| Model (2) | Model (1) |  |
|  | 2.571  0.028 | Intercept |
|  | -0.711  0.033 | mediaTypeMovie |
|  | -0.374  0.031 | mediaTypeMusic Video |
|  | -0.502  0.033 | mediaTypeOVA |
|  | -0.835  0.037 | mediaTypeTV |
|  | -0.381  0.045 | mediaTypeTV Special |
|  | -0.652  0.032 | mediaTypeWeb |
|  | 0.01  0.00035 | duration |
|  | 0.218  0.019 | is\_descriptionYes |
|  | 0.065  0.003 | tags\_count |
|  | -0.076  0.013 | count\_warning |
|  | 0.000002  0.000012 | watched |
|  | 0.000051  0.000018 | votes |
|  | 0.3885 | (6) R-adjusted |
|  | 6830 | (7) N |

Text

Description automatically generatedהפלט המתקבל:

1. הסבר ודיווח המקדמים שהתקבלו:

|  |  |  |
| --- | --- | --- |
|  | מובהקות האומדן | משמעות המקדמים ותוצאות |
| Intercept | מובהק | כאשר כל המשתנים שווים לאפס, הדרוג יהיה שווה ל- 2.571. אין הרבה משמעות לחותך במודל שלנו, כלומר מה זאת אומרת שיש סרט באורך אפס? |
| mediaTypeMovie | מובהק | כאשר יתר המשתנים ללא שינוי, התוספת הממוצעת ל-rating הינה -0.711 עבור האומדן של mediaTypeMovie |
| mediaTypeMusic Video | מובהק | כאשר יתר המשתנים ללא שינוי, התוספת הממוצעת ל-rating הינה -0.374 עבור האומדן של mediaTypeMusic Video |
| mediaTypeOVA | מובהק | כאשר יתר המשתנים ללא שינוי, התוספת הממוצעת ל-rating הינה -0.502 עבור האומדן של mediaTypeOVA |
| mediaTypeTV | מובהק | כאשר יתר המשתנים ללא שינוי, התוספת הממוצעת ל-rating הינה -0.835 עבור האומדן של mediaTypeTV |
| mediaTypeTV Special | מובהק | כאשר יתר המשתנים ללא שינוי, התוספת הממוצעת ל-rating הינה -0.381 עבור האומדן של mediaTypeTV Special |
| mediaTypeWeb | מובהק | כאשר יתר המשתנים ללא שינוי, התוספת הממוצעת ל-rating הינה -0.652 עבור האומדן של mediaTypeWeb |
| is\_descriptionYes | מובהק | כאשר יתר המשתנים ללא שינוי, התוספת הממוצעת ל-rating הינה 0.218 עבור האומדן של is\_descriptionYes |
| duration | מובהק | עלייה של יחידה אחת ב-duration, כאשר יתר המשתנים ללא שינוי, מביאה לעלייה של 0.01 יחידות ב-rating, בממוצע .  לפי מה שניתחנו והצגנו בגרף השלישי(בחלק הקודם), נגיד שכן התוצאות תואמות למה שהוצג בגרף. |
| tags\_count | מובהק | עלייה של יחידה אחת ב-duration, כאשר יתר המשתנים ללא שינוי, מביאה לעלייה של 0.065 יחידות ב-rating, בממוצע .  המקדם חיובי, והיגיוני להגיד שככל שיש לאנימה יותר תגים ישפיע חיובית על הדרוג. |
| count\_warning | מובהק | עלייה של יחידה אחת ב-count\_warning, כאשר יתר המשתנים ללא שינוי, מביאה לעלייה של 0.01 יחידות ב-rating, בממוצע .  המקדם חיובי, התוצאה שהתקבלה אולי היגיונית, אבל אינה מובנת מאליו. |
| watched | לא מובהק | עלייה של יחידה אחת ב-watched, כאשר יתר המשתנים ללא שינוי, מביאה לעלייה של -0.076 יחידות ב-rating, בממוצע .  שיפוע שלילי, התוצאה אינה תואמת למה שהיינו מצפים. היינו חושבים שככל שיש יותם צופים זאת אומרת שהאנימה פופולארי יותר ולכן יהיה לו דרוג יותר גבוה. |
| votes | מובהק | עלייה של יחידה אחת ב-votes, כאשר יתר המשתנים ללא שינוי, מביאה לעלייה של 0.000051 יחידות ב-rating, בממוצע .  מקדם חיובי, התוצאה היגיונית וכמו שהיינו חושבים. |

1. נבדוק אם המשתנה mediaType חשוב ותורם מידע לדירוג של המשתנה המוסבר rating.

נבדוק את ההשערה על ידי בניית מודל חדש בלי המשתנה mediaType, ונבדוק אם התוצאה מובהקת על ידי מבחן F:

מסכנה: עבור רמת מובהקות ששווה ל-0.05, יצא ש-. ולכן לא נדחה את השערת האפס, ונגיד שהמשתנה mediaType כן חשוב ותורם מידע לחיזוי המשתנה rating.

1. נבדוק אם ישנה הפרה של **הנחת ההטרוסקדסטיות לגבי ערך השונות של הטעות:**

, Homoscedasticity

, Hetroscedasticity

נבדוק את ההשערה בעזרת מבחן Breusch-Pagan:

Text

Description automatically generatedקיבלנו , ולכן נדחה את השערת האפס, ונגיד שהנחת שוויון השונויות אינה מתקיימת.