רגרסיה ומודלים לינאריים 52320 תשע"ה 17־2016
עבודת בית 2 08.06.2017

עבודת בית זו מכילה שאלות תיאורטיות ושאלות חישוביות.

משקל כל סעיף בציון נתון ליד הסעיף. מספר הנקודות הכולל הוא 105. הציון המקסימלי בכל מקרה הוא 100.

השאלות הן בדרגת קושי שונה כך שמומלץ לא להתעכב יתר על המידה על שאלה מסויימת.
ניתן לעשות והגיש את העבודה ביחידים או בזוגות.

אנא הקפידו על ההנחיות הבאות:

- הגישו את תשובותיכם במייל. ־ על הפתרונות להיות מודפסים או כתובים **בכתב יד ברור** וסרוקים.
- כתבו את ת.ז. **(לא את השם!)** בראש העמוד הראשון של כל שאלה. אם הגשתם בזוג כתבו את ת.ז. של שני בני הזוג.
- כתבו את הפתרון לכל שאלה בעמוד נפרד. ציינו בבירור את מספר השאלה והסעיפים.
- **תשובה סופית ללא דרך לא תזכה בניקוד כלשהו (ציון 0).**
- כתבו פתרון מלא אך תמציתי לכל שאלה. נמקו כל שלב בפתרונכם אך אין לצרף כיוונים שלא צלחו, פתרונות אלטרנטיביים וכו' .
- אין לצרף דפי טיוטה ־ הגישו רק את הפתרון הסופי והברור ביותר אליו הגעתם עבור כל שאלה.
- בשאלות החישוביות עליכם להסביר בפירוט את ניתוח הנתונים ולצרף את הקוד של הפונקציות שכתבתם. השאלות משתמשות בקבצי נתונים הנמצאים באתר הקורס.
- ניתן להתייעץ עם חברים לגבי החומר הכללי שנלמד, אבל את העבודה עצמה על כל תלמיד (או זוג) לפתור ולכתוב באופן עצמאי. **העתקות יטופלו בחומרה.**
- **משך העבודה: עשרה ימים** .הגישו את העבודה עד ליום ראשון ה־ **18.06.2017** בשעה 59 : 23 במייל

  ל־michal.hataby@mail.huji.ac.il. פתרונות אשר יוגשו מאוחר יותר לא ייבדקו.
- את שאלה 1 יש להגיש בנוסף בקובץ **R** באימייל בודד לmichal.hataby@mail.huji.ac.il עבור כל תלמיד (או זוג). יש לעקוב בקפידה אחרי הוראות נוספות בשאלות.

בהצלחה!

סימונים: נכתוב משתנים בכתיב וקטורי, כאשר $x, y, ..$ הם וקטורי עמודה. x_i מסמן את האיבר ה־i של וקטור x ו־$\bar{x}$ מסמן את הממוצע של וקטור x. עבור שני וקטורים x, y באורך n המכפלה הסקלרית שלהם היא $x^T y = \sum_{i=1}^{n} x_i y_i$.

1. בשאלה זו עליכם לכתוב פונקציות ב־R המחשבות גדלים שונים עבור בעיית רגרסיה מרובה עם נתונים $\mathbf{X}, y$ ומודל רגרסיה $y = \mathbf{X}\beta + \epsilon$, ועם אומד הרבועים הפחותים $\hat{\beta}$ ווקטור התחזיות $\hat{y}$ הרגילים. בשאלה זו בלבד **אין להשתמש** בפונקציות הרגרסיה ב־R (כמו lm) אלא יש לממש את החישובים בעצמכם. עבור סעיפים (א.)־(ד.) יש להגיש קובץ R בלבד בשם: **Quiz 2 - Q1.R** המכיל 4 פונקציות. על הקובץ להיות כתוב בהתאם לתבנית הנתונה בקובץ Home Quiz 2 - Q1 Template.R שנמצא באתר הקורס. בפרט, יש למלא את ת.ז במקום הנדרש (ללא שם). **אין לשנות את שם הפונקציה או את מבנה הקובץ! בבדיקה הקובץ יורץ כפי שנשלח, והציון ייקבע על פי נכונות הפונקציה - פונקציות שיחזירו תוצאות לא נכונות או יחזירו הודעת שגיאה יקבלו ציון 0!**

(א) [6 נק'] כתבו פונקציה המקבלת כקלט מערך דו ממדי של ערכי X ומערך חד ממדי של ערכי y. הפונקציה צריכה להחזיר את וקטור התחזיות ה־$\hat{y}$ המכיל תחזית עבור כל התצפיות. שם הפונקצייה: $my.predictions$. הפלט צריך להיות וקטור עמודה באורך n (כאשר n הוא מספר התצפיות).

(ב) [10 נק'] יהי וקטור השאריות $e = y - \hat{y}$ ונגדיר את וקטור השאריות המתוקננות e^* ע"י $e_i^* = \frac{e_i}{S\sqrt{[I_n - P_X]_{ii}}}$ כאשר S האומד לסטיית התקן, I_n מטריצת היחידה ו־P_X מטריצת ההטלה על תת המרחב הנפרש ע"י עמודות X. כתבו פונקציה המקבלת כקלט מערך דו ממדי של ערכי X ומערך חד ממדי של ערכי y. הפונקצייה צריכה להחזיר את וקטור השאריות המתוקננות e^* עבור כל התצפיות. שם הפונקציה: $my.standardized.residuals$. הפלט צריך להיות וקטור עמודה באורך n.

(ג) [10 נק'] כתבו פונקציה המקבלת כקלט מערך דו ממדי של ערכי X ומערך חד ממדי של ערכי y וכן רמת סמך מבוקשת $1 - \alpha$ ומחשבת רווח סמך לתוחלת μ_i של y_i עבור כל תצפית. הפונקציה צריכה להחזיר שני וקטורים המהווים את קצוות רווח הסמך $[\mu_i^-, \mu_i^+]$ עבור $\mu_i = x_i^T \beta$ ברמת סמך $1 - \alpha$ עבור כל תצפית x_i (המייוצגת על ידי השורה ה־i במטריצת הקלט X). שם הפונקציה: $my.confidence.mu$. הפלט צריך להיות מטריצה עם n שורות ו־2 עמודות כאשר ערכי μ_i^- בעמודה הראשונה וערכי μ_i^+ בעמודה השניה.

(ד) [10 נק'] כתבו פונקצייה המקבלת כקלט מערך דו ממדי של ערכי X ומערך חד ממדי של ערכי y וכן את רמת הסמך המבוקשת $1 - \alpha$. הפונקציה צריכה להחזיר שני מספרים המהווים את קצוות רווח הסמך $[\sigma_-^2, \sigma_+^2]$ ברמת סמך $1 - \alpha$ עבור σ^2. שם הפונקציה: $my.confidence.sigma$

(ה) [15 נק'] השתמשו בפונקציות שכתבתם כדי לנתח את קובץ הנתונים של השכרת האופניים bikes.txt המופיע במודל. תיאור המשתנים מופיע בקובץ bikesReadme.txt. השתמשו במודל מתרגיל 8, שאלה 3, סעיף א. (המודל **כולל חותך**:)

i. ציירו גרף ובו בציר x מופיעות התחזיות $\hat{y}$, ובציר y השאריות המתוקננות e^*. האם נראה שיש קשר בין ערכי $\hat{y}$ לערכי e^*?

ii. ציירו גרף ובו בציר x מופיעות התחזיות $\hat{y}$, ובציר y מופיעים: 1. ערכי y_i הנצפים, 2. שני קוים המתארים את קצות רווח הסמך $[\mu_i^-, \mu_i^+]$ עבור μ_i עבור $1 - \alpha = 0.95$. חשבו את אחוז התצפיות עבורן ערכי y_i נמצאים מחוץ לרווח הסמך $[\mu_i^-, \mu_i^+]$ - האם התוצאה מפתיעה? הסבירו.

iii. חשבו רווח סמך ברמת סמך 95% לשונות σ^2.

2. בשאלה זו עליכם לנתח קובץ נתונים של ההצבעה בבחירות לכנסת ב־ 2015 בישובים שונים בתלות בפרמטרים דמוגרפיים של כל ישוב. עליכם לקרוא את קובץ הנתונים Elections2015_with_covariates.xlsx מהמודל. בקובץ זה כל שורה מכילה ישוב (הקובץ מכיל רק ישובים בינוניים עם $2000 - 500$ נפש). כל עמודה מכילה משתנה דמוגרפי של היישוב, פרט לשתי העמודות הראשונות, המכילות את שם הישוב ומספר המזהה את היישוב (סמל ישוב) ול־10 העמודות האחרונות, המכילות כל אחת את אחוז ההצבעה למפלגה מסויימת בישוב. (הקובץ מכיל רק את 10 המפלגות שעברו את אחוז החסימה בבחירות). עליכם לנתח את אחוזי ההצבעה עבור **מפלגה אחת בלבד** (המשתנה המוסבר) בעזרת **כל הנתונים הדמוגרפיים** (המשתנים המסבירים). המפלגה אותה עליכם לנתח נקבעת על פי ספרת הביקורת של מס. ת.ז. שלכם (אם מגישים בזוג, עליכם לחבר את שתי ספרות הביקורת של שני בני הזוג ולקחת את ספרת האחדות), על פי המפתח הבא (לפי סדר אלפבתי):

0 ־ אמת (המחנה הציוני), 1 ־ ג (יהדות התורה), 2 ־ ודעם (הרשימה המשותפת), 3 ־ טב (הבית היהודי), 4 ־ כ (כולנו), 5 ־ ל (ישראל ביתנו), 6 ־ מחל (הליכוד), 7 ־ מרצ (מרצ), 8 ־ פה (יש עתיד), 9 ־ שס (שס).

(א) [14 נק'] התאימו מודל רגרסיה לינארית מרובה (כולל חותך) לנתונים. כתבו סיכום קצר הדן בתוצאות הניתוח: חשבו אומדים לכל המקדמים. אילו משתנים הם סיגניפיקנטיים (ברמת מובהקות 0.01)? מהו טיב ההתאמה של המודל? יש לצרף פלט רלוונטי (טבלאות, גרפים וכו') לגיבוי מסקנותיכם.

(ב) [5 נק'] המירו את נתוני ההצבעה למפלגה למספר מנדטים (מתוך 120, **אין** לעגל למנדטים שלמים) וחשבו את ערכי מקדמי הרבועים הפחותים וכן R^2 שיתקבלו אם היינו מתאימים את המודל מחדש עבור מספר מנדטים כמשתנה התלוי, **מבלי** להתאים את המודל מחדש.

(ג) [5 נק'] בדקו ברמת מובהקות $\alpha = 0.05$ האם יש השפעה ל(כלל משתני) ארץ המוצא של יהודים בישוב על אחוז ההצבעה למפלגה. הסבירו באיזה מבחן השתמשתם וכיצד ביצעתם אותו.

(ד) [5 נק'] בדקו ברמת מובהקות $\alpha = 0.05$ האם יש אינטקרציה כפלית בין ההשפעה של שנת ייסוד הישוב להשפעה של גודל הישוב על אחוז ההצבעה למפלגה. הסבירו באיזה מבחן השתמשתם וכיצד ביצעתם אותו.

(ה) [5 נק'] בדקו ברמת מובהקות $\alpha = 0.05$ האם יש אינטקרציה כפלית בין ההשפעה של (כלל משתני) גיל התושבים להשפעה של גודל הישוב על אחוז ההצבעה למפלגה. הסבירו באיזה מבחן השתמשתם וכיצד ביצעתם אותו.

(ו) [5 נק'] מצאו ישובים עבורם תחזית המודל אינה הגיונית ־ איך הייתם משפרים אותה?

(ז) [5 נק'] מצאו את שני הישובים בהם תחזית המודל היא הרחוקה ביותר מאחוז ההצבעה בפועל.

(ח) [10 נק'] חלקו את היישובים לקבוצות על פי סמל מחוז, וחשבו בכל סמל מחוז תחזית לאחוז ההצבעה למפלגה בישובים במחוז זה. שימו לב שעליכם לחשב ממוצע משוקלל על פי גדלי הישובים. הגדירו במדוייק מהו הפרמטר שאותו אתם צריכים לאמוד וכיצד חישבתם את התחזית.

הערות: שימו לב כי חלק מן המשתנים הם מספריים וחלק קטגוריים. תוכלו לקבל עוד מידע על המשתנים בקובץ Demographic_parameters.xlsx. תוכלו לבדוק ולהשוות את תוצאותיכם עבור ישובים או מפלגות ספציפיות לנתונים באתר הבא: `http://votes20.gov.il/cityresults` (ניתן להוריד משם גם קובץ המכיל את תוצאות הבחירות עבור כלל היישובים והמפלגות). לחלק מהסעיפים בשאלה זו תתכן יותר מדרך פתרון אחת אפשרית ־ בחרו את הפתרון הנראה לכם ההגיוני ביותר וכתבו אותו בבהירות. צרפו את הקוד שכתבתם כדי לנתח שאלה זו