רגרסיה ומודלים לינאריים 52320 תשע"ה 15־2014
בוחן 2 14.06.2015

בוחן זה הוא בוחן בית המכיל שאלות תיאורטיות ושאלות חישוביות.

משקל כל סעיף בציון הבוחן 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 הבוחן ביחידים או בזוגות.

אנא הקפידו על ההנחיות הבאות:

- הגישו את תשובותיכם בכתב. השתמשו בדפים חלקים או דפי חשבון (משבצות) ־ על הפתרונות להיות מודפסים או כתובים **בכתב יד ברור**
- כתבו את ת.ז. (**לא את השם!**) בראש כל עמוד של דפי תשובותיכם. אם הגשתם בזוג כתבו את ת.ז. של שני בני הזוג.
- כתבו את הפתרון לכל שאלה בעמוד נפרד. ציינו בבירור את מספר השאלה והסעיפים. לבסוף שדכו את כל הדפים של פתרונותיכם.
- **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 הבוחן עצמו על כל תלמיד (או זוג) לפתור ולכתוב באופן עצמאי. **העתקות יטופלו בחומרה.**
- **משך הבוחן: שבוע** . עליכם להגיש את הבוחן עד ליום שני ה־ 22.06.2015 בשעה 8:00 בבוקר לתא ההגשות של דניאל. פתרונות אשר יוגשו מאוחר יותר לא ייבדקו.
- את שאלות 3 ו4 יש להגיש בשני קבצי R נפרדים באימייל בודד לdanielnevo@gmail.com עבור כל תלמיד (או זוג). יש לעקוב 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. עבור נתונים $(x_1,y_1), ..., (x_n, y_n)$ במודל לרגרסיה פשוטה עם חותך: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ נגדיר את שני האומדים הבאים לשיפוע β_1:

- ניצור את המטריצה $X = [1|x]$ (כלומר העמודה הראשונה היא עמודת אחדות, והעמודה השנייה היא ערכי x) , נחשב את אומד הרבועים הפחותים הסטנדרטי עבור רגרסיה מרובה: $\hat{\beta} = [X^T X]^{-1} X^T y$ וניקח את האיבר **השני** בוקטור $\hat{\beta}$ המתקבל כאומד שלנו.
- נחשב את האומד בשני שלבים: בשלב הראשון נעשה רגרסיה לינארית פשוטה רק עם חותך כאשר x הוא המשתנה התלוי עם המודל: $x_i = \gamma_0 + \epsilon'_i$ ונחשב את השאריות במודל זה $e_i^{(x)} = x_i - \hat{x}_i$. בשלב השני נעשה רגרסיה לינארית פשוטה כאשר y הוא המשתנה התלוי והשארית $e^{(x)}$ היא המשתנה הבלתי תלוי, עם המודל $y_i = b_0 + b_1 e_i^{(x)} + \epsilon''_i$ וניקח את אומד הרבועים הפחותים ל־b_1 במודל זה כאומד שלנו.

(א) [10 נק'] הוכיחו כי האומד המתקבל בשיטה הראשונה זהה לאומד הרבועים הפחותים הרגיל לשיפוע עבור רגרסיה לינארית פשוטה.

(ב) [5 נק'] הוכיחו כי גם האומד המתקבל בשיטה השנייה זהה לאומד הרבועים הפחותים הרגיל לשיפוע עבור רגרסיה לינארית פשוטה - כלומר שלושת האומדים המתקבלים זהים.

(ג) [5 נק'] כעת נשתמש בשתי השיטות לעיל לאמידת החותך β_0: ניקח בשיטה הראשונה את האיבר **הראשון** בוקטור $\hat{\beta}$ ובשיטה השניה את אומד הרבועים הפחותים ל־b_0. האם האומד לחותך בשתי השיטות הוא זהה?

2. עבור מודל רגרסיה מרובה: $y = X\beta + \epsilon$ עם $\epsilon \sim N(0, \sigma^2 I)$ יהי $\hat{\beta} = [X^T X]^{-1} X^T y$ אומד הרבועים הפחותים. נגדיר את שגיאת האמידה הרבועית הממוצעת של $\hat{\beta}$ כתוחלת של הנורמה האוקלידית של ההפרש בין וקטור הפרמטרים β לבין האומד שלו $\hat{\beta}$, כלומר $MSE_\beta \equiv E\left[||\hat{\beta} - \beta||^2\right] = E\left[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)\right]$.

(א) [10 נק'] הוכיחו ששגיאת האמידה הרבועית הממוצעת ניתנת על ידי הביטוי: $MSE_\beta = \sigma^2 trace\left((X^T X)^{-1}\right)$.
(תזכורת: עבור מטריצה רבועית A העקבה $(trace)$ מוגדרת כסכום אברי האלכסון: $trace(A) = \sum_i A_{ii}$)

(ב) [5 נק'] כעת הניחו שמבצעים טרנספורמציות ליניאריות על X ועל y : $X' = aX,\ y' = cy$ עבור $a, c \neq 0$ ועושים רגרסיה לינארית מרובה של y' מול X' עבור המודל $y' = X'\beta' + \epsilon'$. כיצד תשתנה שגיאת האמידה הרבועית הממוצעת עבור $\hat{\beta}'$ ביחס לשגיאה בסעיף הקודם עבור $\hat{\beta}$?

3. בשאלה זו עליכם לכתוב פונקצייה ב־R המחשבת את הסטטיסטי של F ולהשתמש בה. יש להגיש קובץ R בלבד בשם: Quiz 2 - Q3.R. על הקובץ להיות כתוב בהתאם לתבנית הנתונה בקובץ Quiz 2 - Q3 - R template.R שנמצא באתר הקורס. בפרט, יש למלא את ת.ז במקום הנדרש (ללא שם). אין לשנות את שם הפונקציה או את מבנה הקובץ.

(א) [10 נק'] כתבו פונקצייה המקבלת כקלט מערך דו ממדי של ערכי X ומערך חד ממדי של ערכי y וכן מערך I עם אינדקסים של המשתנים אותם רוצים לבחון. הפונקצייה צריכה להחזיר את ערך הסטטיסטי F וכן את את ה־$p-value$ המתאים למבחן F עבור השערת האפס: $H_0 : \beta^{(I)} = 0$ (כלומר המקדמים של כל המשתנים הנתונים ע"י האינדקסים מתאפסים).

(ב) [10 נק'] השתמשו בפונקצייה שכתבתם כדי לנתח את קובץ הנתונים של השכרת האופניים bikes.txt. תיאור המשתנים מופיע בקובץ bikesReadme.txt. השתמשו במודל מתרגיל 8, שאלה 3, סעיף א' ובדקו ברמת מובהקות $\alpha = 0.05$ האם יש השפעה לעונת השנה על כמות ההשכרות היומית.

4. בשאלה זו עליכם לכתוב פונקצייה ב־R המחשבת רווח סמך לשונות השגיאה σ^2. יש להגיש קובץ R בלבד בשם: Quiz 2 - Q4.R. על הקובץ להיות כתוב בהתאם לתבנית הנתונה בקובץ Quiz 2 - Q4 - R template.R שנמצא באתר הקורס. בפרט, יש למלא את ת.ז במקום הנדרש (ללא שם) ־ ולציין בסעיף (ב.) באיזה קובץ נתונים אתם משתמשים ומתי הוצג המודל (באיזה תרגול או תרגיל). אין לשנות את שם הפונקציה או את מבנה הקובץ.

(א) [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.

(ב) [10 נק'] השתמשו בפונקצייה שכתבתם כדי לחשב רווח סמך ברמת סמך 95% לשונות עבור מודל שהוצג בתרגילים או בתרגול (וכללו הפניה מתאימה) הכולל לפחות שני משתנים מסבירים (ובנוסף חותך) באחד מקבצי הנתונים איתם אנו עובדים בקורס ונמצאים באתר.

5. 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 . 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) [20 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 0.005)? האם הנחות המודל מתקיימות? מהו טיב ההתאמה של המודל? יש לצרף פלט רלוונטי (טבלאות, גרפים וכו') לגיבוי מסקנותיכם.

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

(ג) [5 נק'] מצאו את שני הישובים בהם תחזית המודל היא הרחוקה ביותר מאחוז ההצבעה בפועל. בדקו האם מדובר בתצפיות חריגות.

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 (מודפס)