רגרסיה ומודלים לינאריים 52307 תשע"ט 2018־19
בוחן בית 26.12.2018

משקל כל סעיף בציון בוחן הבית נתון ליד הסעיף.

השאלות הן בדרגת קושי שונה כך שמומלץ לא להתעכב יתר על המידה על שאלה מסויימת.
יש לעשות ולהגיש את הבוחן ביחידים.

אנא הקפידו על ההנחיות הבאות:

- הגישו את תשובותיכם במודול ־ על הפתרונות להיות מודפסים או סרוקים **בכתב יד ברור.**

  עליכם להסביר בפירוט את ניתוח הנתונים ולצרף את הקוד של הפונקציות שכתבתם וכן פלטים רלוונטיים. השאלות משתמשות בקבצי נתונים הנמצאים באתר הקורס.

- כתבו את ת.ז. בראש כל עמוד של דפי תשובותיכם.
- 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' .
- ניתן להתייעץ עם חברים לגבי החומר הכללי שנלמד, אבל את הבוחן עצמו על כל תלמיד לפתור ולכתוב באופן עצמאי. **העתקות יטופלו בחומרה.**
- **משך הבוחן: כ־3 שבועות** . עליכם להגיש את הבוחן עד ליום שבת ה־ 19.01.2019 בשעה 23:59 דרך המודל. פתרונות אשר יוגשו מאוחר יותר לא ייבדקו.
- יש להגיש קובץ R **בודד** עם פרטי האנליזה **בנוסף** למסמך הפתרון בו מתוארות התוצאות. יש לעקוב אחרי הוראות נוספות בשאלות.

בהצלחה!

1. עליכם לנתח קובץ נתונים של משחקי כדורסל מה־NBA. עליכם לקרוא את קובץ הנתונים $nba\_data\_2011.txt$ המופיע באתר הקורס. בקובץ זה כל שורה מכילה פרק זמן מסויים של משחק. כל עמודה מכילה משתנה המייצג פרק זמן זה, כולל השחקנים ששיחקו בפרק זמן זה (5 מכל קבוצה), הזמן, מספר הנקודות שנקלע על ידי כל קבוצה וכו'. (הקובץ מכיל את כל משחקי עונת $2010-11$). עליכם לנתח את המשחקים עבור **קבוצה אחת בלבד**. הקבוצה אותה עליכם לנתח נקבעת על פי 2 הספרות האחרונות של מס. ת.ז. שלכם , על פי המפתח בטבלה בעמוד הבא (לפי סדר אלפבתי). מומלץ מאוד להיעזר בקובץ ה־ $template$ בשם $Q2\_nba\_analysis\_template.R$ הנמצא באתר הקורס הקורא את הנתונים, מבצע עיבוד מוקדם (עבור הקבוצה הראשונה ATL) ומחשב משתני עזר דרושים (בפרט, נוצר משתנה בינארי עבור כל שחקן המעיד האם השחקן שיחק בכל פרק זמן והוסרו שחקנים ששיחקו פחות מ־20 אחוז מהדקות בממוצע).

$Num.$	$Team$	$Num.$	$Team$	$Num.$	$Team$
$00-02$	ATL	$30-32$	IND	$60-63$	OKC
$03-05$	BOS	$33-35$	LAC	$64-67$	ORL
$06-08$	CHA	$36-38$	LAL	$68-71$	PHI
$09-11$	CHI	$39-41$	MEM	$72-75$	PHX
$12-14$	CLE	$42-44$	MIA	$76-79$	POR
$15-17$	DAL	$45-47$	MIL	$80-83$	SAC
$18-20$	DEN	$48-50$	MIN	$84-87$	SAS
$21-23$	DET	$51-53$	NJN	$88-91$	TOR
$24-26$	GSW	$54-56$	NOH	$92-95$	UTA
$27-29$	HOU	$57-59$	NYK	$96-99$	WAS

(א) [8 נק'] חשבו מתוך הנתונים את המשתנים המסבירים הבאים בהם נרצה להשתמש במודל רגרסיה : 1. $Minutes$ ־ מספר הדקות שעברו מתחילת המשחק (ניתן לחישוב מהמשתנה $StartTime$ ע"י לקיחת ספרות הדקות או מהמשתנה $ElapsedSecs$ ע"י חישוב סכום מצטבר). 2. $Days$ ־ מספר הימים שעברו מהמשחק הראשון (ניתן לחישוב מהמשתנה $GameID$ כאשר 8 הספרות הראשונות בו הן בפורמט $Year:MM:DD$). שמרו אותם בקובץ $MinutesDays.RData$ באמצעות הפקודה $save$ וצרפו קובץ זה לפתרונכם.

(ב) [15 נק'] התאימו מודל רגרסיה לינארית מרובה (עם חותך) לנתונים עבור המשתנה המוסבר: נקודות לדקה ($PointsPerMinute$) מול המשתנים המסבירים הבאים:

שמות כל השחקנים, (יש להשתמש בכל המשתנים מעמודה 56 ואילך פרט לאחרונה) בית/חוץ (במשתנה $Home$), מספר הדקות שעברו מתחילת המשחק (חושב בסעיף קודם) , ומספר הימים שעברו מהמשחק הראשון (חושב בסעיף קודם).

כתבו סיכום קצר הדן בתוצאות הניתוח: חשבו אומדים לכל המקדמים. אילו משתנים הם סיגניפיקנטיים (השתמשו בסעיף זה ובסעיפים הבאים ברמת מובהקות 0.01)? מהו טיב ההתאמה של המודל? יש לצרף פלטים רלוונטיים (טבלאות, גרפים וכו') לגיבוי מסקנותיכם.

(ג) [12 נק'] בדקו האם הנחות המודל על השגיאות מתקיימות (1. תוחלת אפס, 2. שונות שווה, 3. נורמליות) ־ הסבירו וצרפו פלטים רלוונטיים.

(ד) [6 נק'] בדקו האם למשתני השחקנים (כקבוצה) יש תרומה מובהקת למשתנה המוסבר. הסבירו במדוייק באיזו שיטה השתמשתם ומהן מסקנותיכם.

(ה) [6 נק'] בדקו עבור כל שחקן בנפרד האם התרומה שלו למשתנה המוסבר (נקודות לדקה) במשחקי הבית שונה באופן מובהק מתרומתו במשחקי החוץ. הסבירו במדוייק באיזו שיטה השתמשתם ומהן מסקנותיכם.

(ו) [7 נק'] בדקו עבור כל **זוג** שחקנים בנפרד האם יש ביניהם אינטרקציה מובהקת בתרומתם למשתנה המוסבר (ביחס למודל הלינארי מסעיף (ב.) שהוא ללא אינטרקציות). הסבירו במדוייק באיזו שיטה השתמשתם ומהן מסקנותיכם.

(ז) [6 נק'] חשבו את ההנפה ($leverage$) של כל תצפית ובדקו אם יש תצפיות חריגות פוטנציאליות על פי מדד זה עם הנפה גדולה מ־$\frac{4p}{n}$

(ח) [6 נק'] חשבו את מרחק קוק $(Cook-distance)$ של כל תצפית ובדקו אם יש תצפיות חריגות בפועל על פי מדד זה עם מרחק קוק הגדול לפחות פי 3 ממרחק הקוק הממוצע על פני כל התצפיות.

(ט) [10 נק'] חשבו עבור כל תצפית רווח סמך $[\mu_i^-,\mu_i^+]$ ברמת סמך $1-\alpha=0.95$ לתוחלת μ_i של y_i. ציירו גרף ובו בציר x מופיעות התחזיות $\hat{y}$, ובציר y מופיעים: 1. ערכי y הנצפים, 2. שני קוים המתארים את קצות רווח הסמך $[\mu_i^-,\mu_i^+]$ עבור $1-\alpha=0.95$.

(י) [10 נק'] חשבו עבור כל תצפית רווח חיזוי $[y_i^-,y_i^+]$ כך שמתקיים: $P(y_i\in[y_i^-,y_i^+])=1-\alpha=0.95$ עבור מדגם חדש: כלומר תצפיות (x_i,y_i) שיתקבלו עבור ערכי $\boldsymbol{X}$מהמדגם הנתון וערכי y חדשים ובלתי תלויים בערכי ה־y של המדגם הנתון. הסבירו את הקשר לרווח הסמך מהסעיף הקודם. ציירו גרף ובו בציר x מופיעות התחזיות $\hat{y}$, ובציר y מופיעים: 1. ערכי y הנצפים, ו־2. שני קוים המתארים את קצות הרווחים $[y_i^-,y_i^+]$ עבור $1-\alpha=0.95$.

(יא) [14 נק'] בסעיף זה חשבו משתנה מסביר נוסף שהוא אורך פרק הזמן בדקות t_i עבור תצפית i (ניתן לחישוב מהמשתנה $ElapsedSecs$). למספר הנקודות הנקלע בפרק זמן מסויים יש תוחלת ושונות הפרופורציוניות לאורך פרק הזמן t_i. לכן עבור המשתנה המוסבר נקודות לדקה ($PointsPerMinute$) נצפה לשונות שהיא פרופורציונית ל־$\frac{1}{t_i}$ (עבור פרקי זמן קצרים יותר השונות גבוהה יותר). כדי לפצות על כך, נתאים בסעיף זה מודל של $weighted-least-squares$ כאשר לכל תצפית ניתן משקל של $w_i=t_i$ (ניתן לעשות זאת ב־R באמצעות הוספת $'weights'$ בפונקציה $lm()$). במקרה זה אומד הריבועים הפחותים מוחלף באומד $\hat{\beta^{(w)}}=[\boldsymbol{X^T W X}]^{-1}\boldsymbol{W X^T} y$ כאשר $\boldsymbol{W}=diag(w_1,...,w_n)$ מטריצה אלכסונית.

(i) חשבו את אומד ה־$weighted-least-squares$ על פני כל הנתונים והשוו את האומדים שקיבלתם לאומדי הריבועים הפחותים. האם המשתנים הסיגניפיקנטיים השתנו? האם סטיות התקן החדשות קטנות/גדולות יותר לעומת אומדי הריבועים הפחותים?

(ii) חלקו את התצפיות לשתי קבוצות שוות גודל: קבוצה אחת (ה־$train$) בחצי העונה הראשונה וקבוצה שניה (ה־$test$) בחצי העונה השנייה על פי מספר הימים שעברו מהמשחק הראשון מהסעיף הראשון. התאימו את אומד הרבועים הפחותים ואת אומד ה־ $weighted-least-squares$ עבור ה־$train$ והשתמשו בשני המודלים שנאמדו כדי לחשב תחזיות $\hat{y}_i,\hat{y_i^{(w)}}$עבור ה־$test$. השוו את ה־$SSE$ המתקבל על ה־$test$ בין שני המודלים ־ איזה מודל משיג שגיאה נמוכה יותר? הסבירו.

הערות: שימו לב כי חלק מן המשתנים הם מספריים וחלק קטגוריים. תוכלו לקבל עוד מידע על המשתנים בקובץ $NBA\_header.txt$ במודל. לחלק מהסעיפים תתכן יותר מדרך פתרון אחת אפשרית ־ בחרו את הפתרון הנראה לכם ההגיוני ביותר וכתבו אותו בבהירות. צרפו את הקוד שכתבתם כדי לנתח את הנתונים .