

רגרסיה ומודלים לינאריים 52320 תשע"ו 16-2015

בוחן 1 22.12.2015

משקל כל סעיף הוא 20 נקודות כך שמספר הנקודות הכולל הוא 120. בכל מקרה, ציון הבוחן הוא 100 לכל היותר. שימו לב שהשאלות הן בדרגת קושי שונה כך שמומלץ לא להתעכב יתר על המידה על שאלה מסויימת. אנא הקפידו על ההנחיות הבאות:

- כתבו את ת.ז. (לא את השם!) בראש כל עמוד של טופס הבחינה.
- אין לצרף לטופס דפים נוספים.
- אין לתלוש דפים מטופס הבחינה.
- לתשומת לבכם לגבי השאלות הפתוחות:
- תשובה סופית ללא דרך לא תזכה בניקוד כלשהו (ציון 0).
- בשאלות הפתוחות יש לכתוב את הפתרון רק במקום המוקצה לכך, מעל לכל קו כתבו שורה אחת בלבד בכתב יד קריא. (השאלות נכתבו כך שניתן לכתוב פתרון תמציתי לכל סעיף).
- מגבלת המקום תאכף באופן קפדני. פתרונות אשר יחרגו מהמקום המותר, יהיו בכתב קטן מכדי שיהיה קריא, ו/או יכללו יותר משורת כתב אחת לכל קו לא ייבדקו.
- מומלץ מאוד לפתור תחילה את השאלה במחברת הטייטה ולהעתיק את עיקר הפתרון אל הטופס רק לאחר בדיקה. חומר עזר מותר: מחשבון.

משך הבוחן: שעה

בהצלחה!

סימונים: נכתוב משתנים בכתוב וקטורי, כאשר x, y, \dots הם וקטורי עמודה. x_i מסמן את האיבר ה- i של וקטור x . עבור שני וקטורים x, y באורך n המכפלה הסקלרית שלהם היא $x^T y = \sum_{i=1}^n x_i y_i$. תזכורת: עבור מ"מ וקטורי z ומטריצה A מתקיים: $E(Az) = AE(z)$; $Var(Az) = AVar(z)A^T$.

1. 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 $MSE_{\hat{\beta}} \equiv E[\|\hat{\beta} - \beta\|^2] = E[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)]$.

(א) הוכיחו ששגיאת האמידה הרבועית הממוצעת ניתנת על ידי הביטוי: $MSE_{\hat{\beta}} = \sigma^2 \text{trace}((X^T X)^{-1})$. (תזכורת: עבור מטריצה רבועית A העקבה (trace) מוגדרת כסכום אברי האלכסון: $\text{trace}(A) = \sum_i A_{ii}$)

(ב) כעת הניחו שמבצעים טרנספורמציות לינאריות על X ועל y : $X' = aX + b\mathbf{1}\mathbf{1}^T$, $y' = cy + d\mathbf{1}$. עבור סקלרים a, b, c, d עם $a, c \neq 0$ וכאשר $\mathbf{1}$ וקטור עמודה ועושים רגרסיה לינארית מרובה של y' מול X' עבור המודל $y' = X'\beta' + \epsilon'$. כיצד תשתנה שגיאת האמידה הרבועית הממוצעת של האומד $\hat{\beta}'$ עבור β' ביחס לשגיאה בסעיף הקודם של האומד $\hat{\beta}$ עבור β ?

2. מעוניינים להתאים מודל רגרסיה עבור נתוני תמותה ב-60 ערים בארה"ב (הנתונים נאספו בשנת 1960).

המשתנה המוסבר הינו mortrate והוא מציין שיעורי תמותה ל-100,000 איש בשנה. לדוגמא בעיר בוסטון נרשמו 934.7 מקרי מוות עבור כל 100,000 איש.

המשתנים המסבירים שנאספו על כל עיר הם: כמות משקעים שנתיים בס"מ (precip), טמפרטורה ממוצעת בינואר בצלסיוס (jantemp), חציון מספר שנות חינוך (educ), אחוז לא-לבנים בעיר, בטווח בין אפס למאה (pctnonwt). המודל כולל גם חותך.

המטריצה $(X^T X)^{-1}$ חושבה והתקבל

	intercept	precip	jantemp	educ	pctnonwt
intercept	5.3935688384	-8.900133e-04	1.305460e-02	-4.100702e-01	-3.936508e-03
precip	-0.0008900133	4.002277e-07	1.471286e-07	5.052542e-05	-3.744107e-06
jantemp	0.0130546034	1.471286e-07	7.104761e-04	-1.029001e-03	-2.262434e-04
educ	-0.4100701597	5.052542e-05	-1.029001e-03	3.272483e-02	3.495447e-04
pctnonwt	-0.0039365084	-3.744107e-06	-2.262434e-04	3.495447e-04	3.288717e-04

(תזכורת: עבור הסימון שמתקבל בפלט: לדוגמא $(3.21e - 04 = 0.000321)$)

לפניכם פלט תוצאות רגרסיה כפי שהתקבל ב R. מבחן ה-F המתואר בפלט משווה בין המודל המלא (כלומר $\beta \in \mathbb{R}^5$) לבין מודל המכיל רק את החותך, כלומר: $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

Call: `lm(formula = mortrate ~ precip + jantemp + educ + pctnonwt, data = airpol)`

Residuals:

Min	1Q	Median	3Q	Max
-91.353	-25.281	-2.066	26.452	80.879

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1077.25383	88.59183	12.160	< 2e-16 ***
precip	0.03070	0.02413	1.272	0.20870
jantemp	-3.50989	1.01679	-3.452	0.00108 **
educ	-19.90855	A	B	0.00558 **
pctnonwt	4.74294	0.69178	6.856	6.49e-09 ***

--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.15 on 55 degrees of freedom

Multiple R-squared: 0.6494, Adjusted R-squared:

F-statistic: C on 4 and 55 DF, p-value: 5.705e-12

(א) בכיתה למדנו על מבחן F כללי עבור השערת האפס: $H_0 : A\beta = a$ כאשר $A \in \mathbb{R}_{(p-k) \times p}$ מטריצה ו- $a \in \mathbb{R}^{p-k}$ וקטור. כתבו את המטריצה A והוקטור a המתאימים למבחן לעיל.

(ב) כתבו ביטוי לסטטיסטי של F עבור המבחן לעיל (כלומר מודל עם כל המשתנים לעומת מודל בו יש רק חותך) במקרה הכללי כפונקציה של n, p, SSE, SST (כאן p מספר המשתנים כולל החותך). תזכורת: עבור מבחן F הכללי מהסעיף

הקודם, ניתן לכתוב את הסטטיסטי F כ- $F = \frac{(\|y^{(0)} - y\|^2 - \|\hat{y} - y\|^2)/(p-k)}{\|\hat{y} - y\|^2/(n-p)}$ כאשר \hat{y} וקטור התחזיות במודל הלא מוגבל $y^{(0)}$ הוא וקטור התחזיות במודל תחת H_0 .

(ג) כתבו ליד כל אות את המספר החסר (יש לדייק עד 2 ספרות אחרי הנקודה. אין צורך בכתיבת הסבר):

----A

----B

----C

(ד) השתמשו ב- $\hat{\beta}$ כדי לחשב אומדן לתוחלת שיעור התמותה ל-100,000 איש בשנה בעיר בה הטמפרטורה הממוצעת בינואר היא 5 מעלות, כמות המשקעים הממוצעת בשנה במ"מ היא 1150 מספר שנות ההשכלה החציוני הינו 10.5 ואחוז הלא לבנים בעיר הוא 12.
