אוניברסיטת בן גוריון
המחלקה להנדסת תעשייה וניהול
קורס מודלים של רגרסיה לינארית
364-1-1061

מבחן מועד א' סמסטר ב' – תשע"ז

12.07.2017

פרופ' הלל בר גרא ופרופ' ישראל פרמט

משך המבחן: שלוש שעות
חומר עזר מותר: 5 דפי נוסחאות, טבלאות סטטיסטיות ומחשבון
הנחיות חשובות:
- יש לפתור את כל השאלות
- יש ללוות את דרך הפתרון בהסברים מילוליים קצרים
- ניתן להשתמש בכל הנוסחאות והתוצאות שפותחו או הוצגו בכיתה ואין צורך להוכיחן מחדש אלא אם כן זה נדרש במפורש
- כאשר לא מצוינת ר"מ נדרשת הניחו ר"מ של 5%
-בכל מבחן השערות יש לנסח את ההשערות הנבדקות, לציין מהו סטטיסטי המבחן וכיצד הוא מתפלג, ולכתוב מהו אזור הדחייה (זאת כמובן בנוסף לתוצאת המבחן)

בהצלחה ☺

שאלה 1 (40 נקודות)

חוקר האופניים מייק בייק רצה לבדוק את הגורמים המשפיעים על כמות האופניים המושכרים ביום. לשם כך אספו נתונים במשך שנה, כלומר 365 תצפיות, על כמות האופניים המושכרים באחת הערים בארה"ב. בנוסף לכמות האופניים, רשמו החוקרים עבור כל יום את הטמפ' (temp, בצלסיוס) הממוצעת באותו יום, את מהירות הרוח (windspeed) הממוצעת (במטר בשנייה) ואת אחוזי הלחות (humidity). החוקרים החליטו לנרמל את הנתונים , על ידי חלוקה בערך המקסימלי עבור כל משתנה רציף.

נתונות המטריצות הבאות:

$$X^tX = \begin{pmatrix} 365 & 177.6327 & 234.9376 & 69.8620 \\ 177.6327 & 99.5321 & 115.8323 & 33.3934 \\ 234.9376 & 115.8323 & 159.2745 & 44.0696 \\ 69.8620 & 33.3934 & 44.0696 & 15.5238 \end{pmatrix};$$

$$(X^tX)^{-1} = \begin{pmatrix} 0.1020 & -0.0333 & -0.0888 & -0.1354 \\ -0.0333 & 0.0787 & -0.0127 & 0.0168 \\ -0.0888 & -0.0127 & 0.1323 & 0.0516 \\ -0.1354 & 0.0168 & 0.0516 & 0.4910 \end{pmatrix}$$

החוקרים הריצו מודל רגרסיה לינארית וקיבלו את הפלט הבא:

```
Call:
lm(formula = cnt ~ temp + hum + windspeed)

Residuals:
     Min       1Q   Median       3Q      Max 
-2899.81  -507.27   -27.72   646.38  1954.90 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2401.6      259.8   9.244  ******
temp          5590.1      228.1       A  ******
hum          -1469.9      295.8  -4.969  ******
windspeed    -4024.2          B       C  ******
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: **** on **** degrees of freedom
Multiple R-squared:  0.6549,	Adjusted R-squared:  ****
F-statistic: **** on 3 and 361 DF,  p-value: ****
```

א. (8 נקודות) חשבו את הערכים המספריים של A, B, C בפלט.

$$A = t_{st}(\widehat{\beta_0}) = \frac{\widehat{\beta_0}}{\sqrt{\hat{V}(\widehat{\beta_0})}} = \frac{5590.1}{228.1} = 24.507$$

$$B = \sqrt{\hat{V}(\widehat{\beta_3})} = \sqrt{MSE \cdot (X^tX)^{-1}_{4,4}}$$

$$\sqrt{\hat{V}(\widehat{\beta_2})} = 295.8 = \sqrt{MSE \cdot (X^tX)^{-1}_{3,3}} = \sqrt{MSE \cdot 0.1323} \rightarrow MSE = 661357.8231$$

$$B = \sqrt{\hat{V}(\widehat{\beta_3})} = \sqrt{MSE \cdot (X^tX)^{-1}_{4,4}} = \sqrt{661357.8231 \cdot 0.491} = 569.848$$

$$C = t_{st}(\widehat{\beta_3}) = \frac{\widehat{\beta_3}}{\sqrt{\hat{V}(\widehat{\beta_3})}} = -\frac{4024.2}{569.848} = -7.061$$

ב. (8 נקודות) ידוע כי הערך המקסימלי של הטמפ' (לפני הנירמול) הוא 41, של הלחות (humidity) 100 ושל מהירות הרוח- 67. חשבו את האומדנים למקדמי הרגרסיה עבור נתונים לא מנורמלים.

האומדן של כל פרמטר יקטן פי הערך לפיו נורמל כל משתנה

$$\underline{\hat{\beta}} = (X^tX)^{-1}X^tY \rightarrow \widehat{\beta_1} = \frac{5590.1}{41} = 136.343, \widehat{\beta_2} = \frac{-1469.9}{100} = -14.699, \widehat{\beta_3} = \frac{-4024.2}{67} = -60.062$$

ג. (8 נקודות) בנו טבלת ANOVA עבור המודל וחשבו את מדד טיב ההתאמה. על איזו השערה ניתן לענות באמצעות הטבלה? מהי המסקנה?

על פי הפלט-

$$R^2 = 0.6549 = 1 - \frac{SSE}{SST}$$

נחשב את ה-MSE לפי אומדני השונות של האומדים לביטות, תוך שימוש במטריצה $(X^tX)^{-1}$

$$\widehat{cov}\left(\underline{\hat{\beta}}\right) = MSE \cdot (X^tX)^{-1} \rightarrow \hat{V}(\hat{\beta}_i) = MSE \cdot (X^tX)^{-1}{}_{i+1,i+1}$$

$$\hat{V}(\widehat{\beta_0}) = MSE \cdot 0.1020 = 259.8^2 \rightarrow MSE = 661725.8824 \rightarrow SSE = MSE(n-p)$$

$$= 661725.8824 \cdot 361 = 238883043.5 \rightarrow SST = \frac{SSE}{(1-0.6549)}$$

$$= \frac{238883043.5}{0.3451} = 692213977.2$$

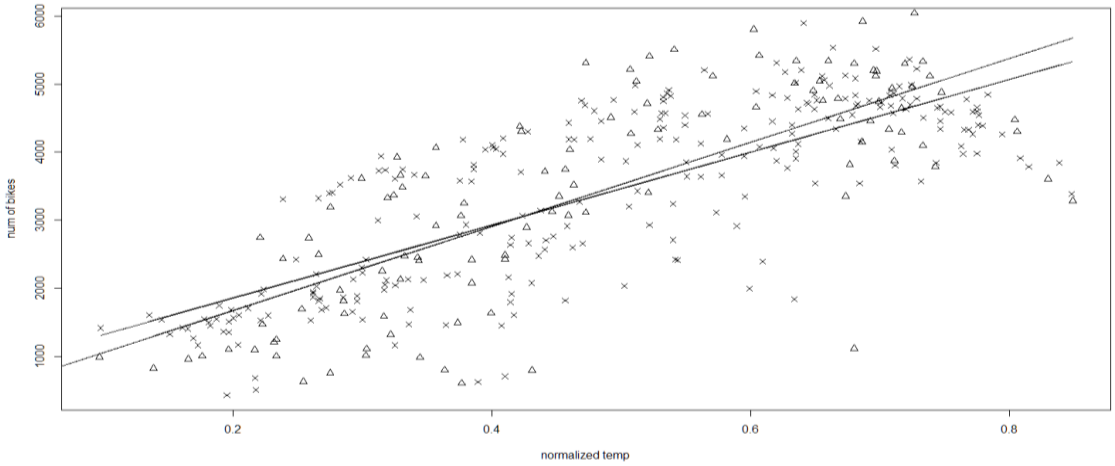
$Source$	SS	$d.f$	MS	F
R	453330933.7	3	151110311.2	$F_{st} = \frac{151110311.2}{661725.8824} = 228.357$
E	238883043.5	361	661725.8824	
T	692213977.2	364		

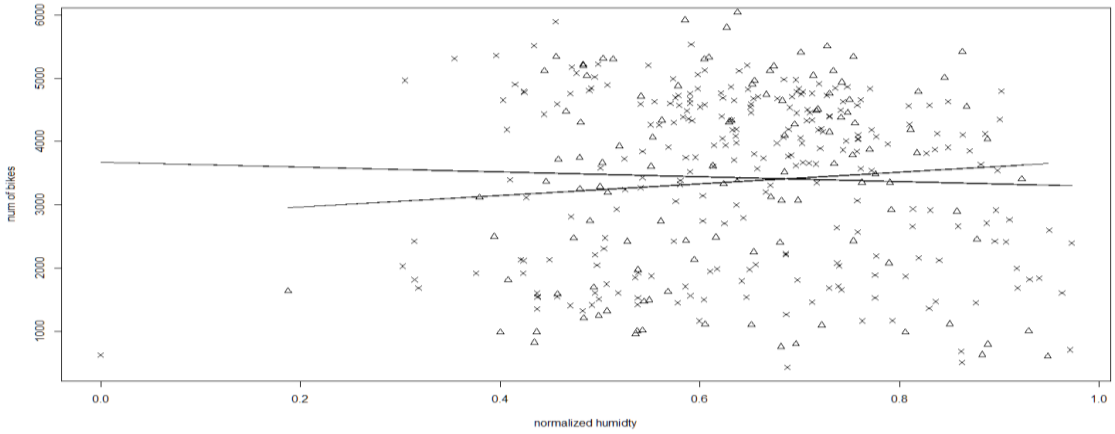
ההשערה עליה ניתן לענות באמצעות הטבלה היא האם אחד המשתנים המסבירים מסביר את המשתנה המוסבר, או במילים אחרות, האם כל מקדמי הרגרסיה שווים ל-0, או אחרת. המסקנה היא שהמודל מובהק, כלומר לפחות אחד המשתנים המסבירים אכן מסביר את המוסבר.

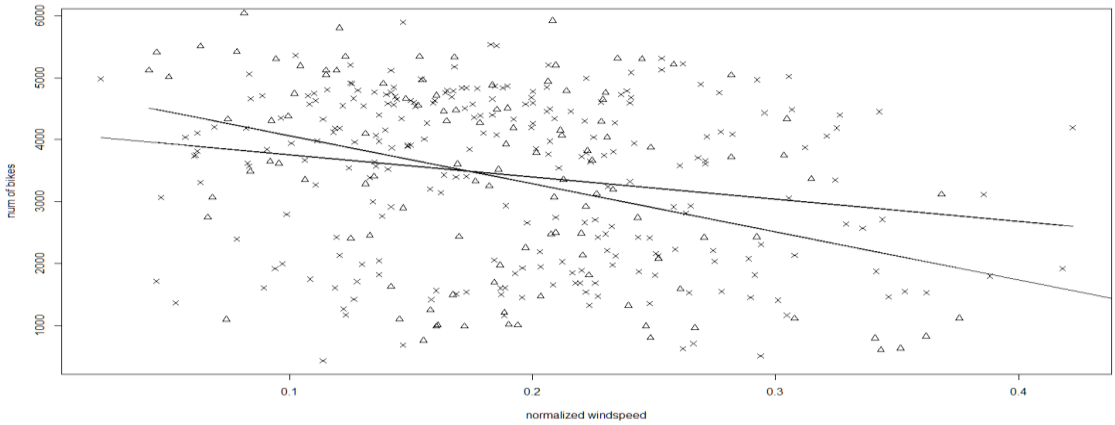
טיב התאמה-

ד. $R^2_{adj} = 1 - \frac{SSE/_{n-p}}{SST/_{n-1}} = 1 - \frac{661725.8824}{1901686.751} = 0.652$

ה. (8 נקודות) החוקרים לא היו מסופקים והחליטו לבדוק עבור כל יום האם זה היה יום עבודה או יום חופש. הסבירו באופן תיאורי את המשמעות המעשית של האינטראקציות המוצגות בתרשימים הבאים (אין צורך להתייחס למובהקות הסטטיסטית).







בתרשים הראשון נראה שלטמפ' מתאם חיובי גבוה עם כמות האופניים המושכרים ובימי חול הקשר "חזק" יותר, כלומר שביום עבודה ככל שהטמפ' עולה, כך עולה כמות האופניים המושכרים, בקצב גבוה יותר מאשר ביום חופש. אך ככל הנראה, אין לאינטראקציה השפעה על המשתנה המוסבר.

בתרשים השני קיים הבדל קטן בין שיפועי הקווים, אך עם זאת אחד מהשיפועים נראה ממש מקביל לציר ה-X, מה שמעיד באחת מהקטגוריות אין השפעה של הלחות על כמות האופניים המושכרים ובקטגוריה השנייה ישנה השפעה מועטה. נראה שעל סמך התרשים עיקר ההבדל הוא באפקט הראשי ולא

באינטראקציה, היות והלחות אינה משפיעה בצורה שונה משמעותית בימי עבודה אל מול ימי חופש. משמעות האינטראקציה- ביום עבודה ישנה ירידה בכמות האופניים המושכרים ככל שהלחות עולה לעומת יום חופש בו הכמות עולה ככל שהלחות עולה.

בתרשים השלישי נראה שיש הבדל בין שיפועי הקווים, ולכן ניתן להגיד שלמהירות הרוח השפעה שונה בימי עבודה אל מול ימי חופש, כלומר שהשפעת האינטראקציה ניכרת, ומשמעותה שביום עבודה, כמות האופניים המושכרים קטנה ככל שמהירות הרוח עולה, בקצב נמוך יותר מאשר ביום חופש.

החוקרים טוענים שגם לעונה (חורף winter, קיץ summer, סתיו fall, אביב spring) ישנה השפעה על כמות האופניים. לאחר בחירת המשתנים התקבל המודל הבא

```
Call:
lm(formula = cnt ~ temp + season + workingday + windspeed + hum +
    temp:season + temp:workingday + season:windspeed + workingday:wind
speed +
    season:hum)

Residuals:
     Min       1Q   Median       3Q      Max
-2559.84  -284.83    66.45   344.50  1811.70

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             1292.5      388.6   3.326 0.000976 ***
temp                    5275.7      728.7   7.240 2.92e-12 ***
seasonSummer             726.2      559.2   1.299 0.194957
seasonFall              8319.7      941.2   8.840  < 2e-16 ***
seasonWinter            3977.9      532.3   7.474 6.44e-13 ***
workingday1              234.6      263.3   0.891 0.373612
windspeed              -2991.7     1002.8  -2.983 0.003055 **
hum                    -1055.9      413.1  -2.556 0.011024 *
temp:seasonSummer       1502.8      818.2   1.837 0.067127 .
temp:seasonFall        -7483.1     1151.5  -6.499 2.82e-10 ***
temp:seasonWinter       1966.5      949.3   2.071 0.039054 *
temp:workingday1       -1087.2      355.2  -3.061 0.002380 **
seasonSummer:windspeed  -798.8     1226.1  -0.652 0.515151
seasonFall:windspeed   -1977.4     1344.6  -1.471 0.142288
seasonWinter:windspeed -2645.8     1116.8  -2.369 0.018387 *
workingday1:windspeed   1593.3      879.3   1.812 0.070842 .
seasonSummer:hum        -593.5      585.4  -1.014 0.311359
seasonFall:hum         -2887.2      673.0  -4.290 2.32e-05 ***
seasonWinter:hum       -4363.6      686.5  -6.357 6.51e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 584.7 on 346 degrees of freedom
Multiple R-squared:  0.829,    Adjusted R-squared:  0.8201
F-statistic: 93.21 on 18 and 346 DF,  p-value: < 2.2e-16
```

ו. (8 נקודות) ענו על המשפטים הבאים בנכון, לא נכון או אין מספיק מידע בשביל לדעת. נמקו תשובותיכם.

- ללחות השפעה שונה בעונות שונות על כמות האופניים המושכרים

נכון. שתיים מהאינטראקציות של משתנה הלחות עם משתני הדמה של עונות השנה מובהקות, ולכן ללחות ישנה השפעה שונה בכל אחת מעונות השנה על כמות האופניים המושכרים.

- אין הבדל בין תוחלת כמות האופניים המושכרים בקיץ ובאביב

נכון. קבוצת הבסיס היא האביב ומשתנה הדמי עבור הקטגוריה "קיץ" אינו מובהק.

ניתן גם להגיד שלא ניתן לדעת, היות והמשתנה עונה נמצא באינטרקאציה עם משתנים אחרים- למרות שכלל האינטראקציות אינן מובהקות.

- בימי עבודה (workingday), ככל שהטמפ' גבוהה יותר כך שוכרים פחות אופניים

לא נכון. מקדם האינטראקציה בין משתנה הטמפ' ליום עבודה אומנם שלילי, -1087.2, אך הוא מעיד על ההבדל בקצב השינוי לעומת קבוצת הבסיס- ימי החופש. קצב השינוי של כמות האופניים הנמכרים כפונק' של הטמפ' הינו 5275.7, ולכן על כל עלייה ביח' טמפ' ביום עבודה, תגדל תוחלת כמות האופניים הנמכרים בגודל חיובי.

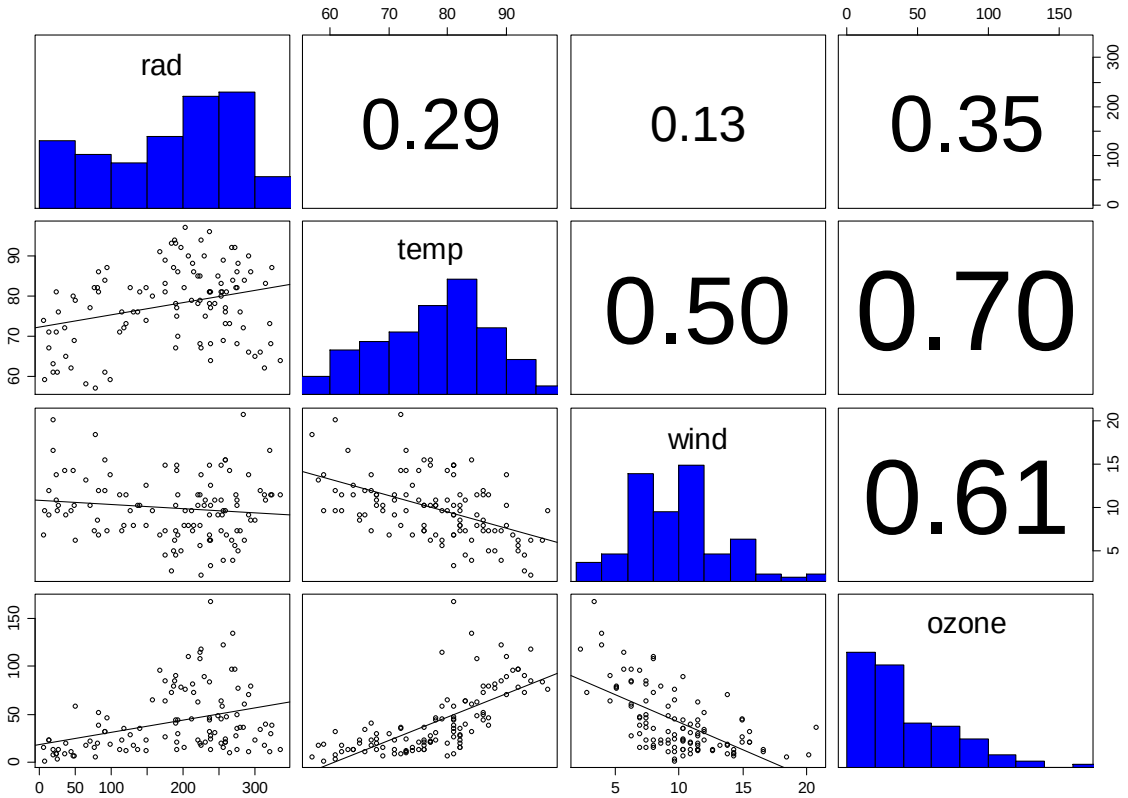
- מהירות הרוח אינה משפיעה על כמות האופניים המושכרים בקיץ.

לא נכון. אומנם משתנה האינטראקציה אינו מובהק, אך משתנה מהירות הרוח בקבוצת הבסיס מובהק, ולכן ההשפעה בקיץ אומנם אינה שונה מההשפעה באביב, אך היא קיימת ומובהקת.

שאלה מספר 2 (40 נקודות)

חוקרים רצו לבדוק האם בין ריכוז האוזון באוויר (ozone) קשור למהירות הרוח (wind), טמפרטורת האוויר (temp) ועוצמת קרינת שמש (rad).

א. (6 נקודות) בשרטוט הבא ניתן לראות את תרשימי הפיזור והמתאמים בין המנבאים ולבין התלוי והמנבאים בינם לבין עצמם. מה ניתן ללמוד משרטוט זה? איזה מנבא הייתם בוחרים להכניס ראשון למודל לפי שיטת "צעדים קדימה"? נמק.



נראה שבין הרוח לבין האוזון ובין הטמפ' לאוזון ישנו קשר לא לינארי לחלוטין. ניתן לראות זאת על ידי ריכוז תצפיות במרכז התחום מתחת לקו הרגרסיה הלינארית, ובקצוות תצפיות מעל לקו הרגרסיה הלינארית. כמו כן נראה שאין קשר (פיזור אקראי) בין עוצמת קרינת השמש לבין ריכוז האוזון באוויר- והמתאם בין המשתנים מחזק את זה.

נראה שיש קשר שלילי בין משתנה הרוח לבין משתנה הטמפ'.

המנבא בעל המתאם הגבוה ביותר הוא הטמפ' ולכן הוא ייכנס ראשון בשיטת רגרסיה לפנים.

ב. (7 נקודות) החוקרים התאימו את המודל הרגרסיה הבא:

```
ozone ~ rad + I(rad^2) + temp + I(temp^2) + wind + I(wind^2)

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  2.985e+02  1.014e+02   2.942  0.00402
rad          1.349e-01  8.795e-02   1.533  0.12820
I(rad^2)    -2.052e-04  2.546e-04  -0.806  0.42213
temp        -6.584e+00  2.738e+00  -2.405  0.01794
I(temp^2)    5.221e-02  1.783e-02   2.928  0.00419
```

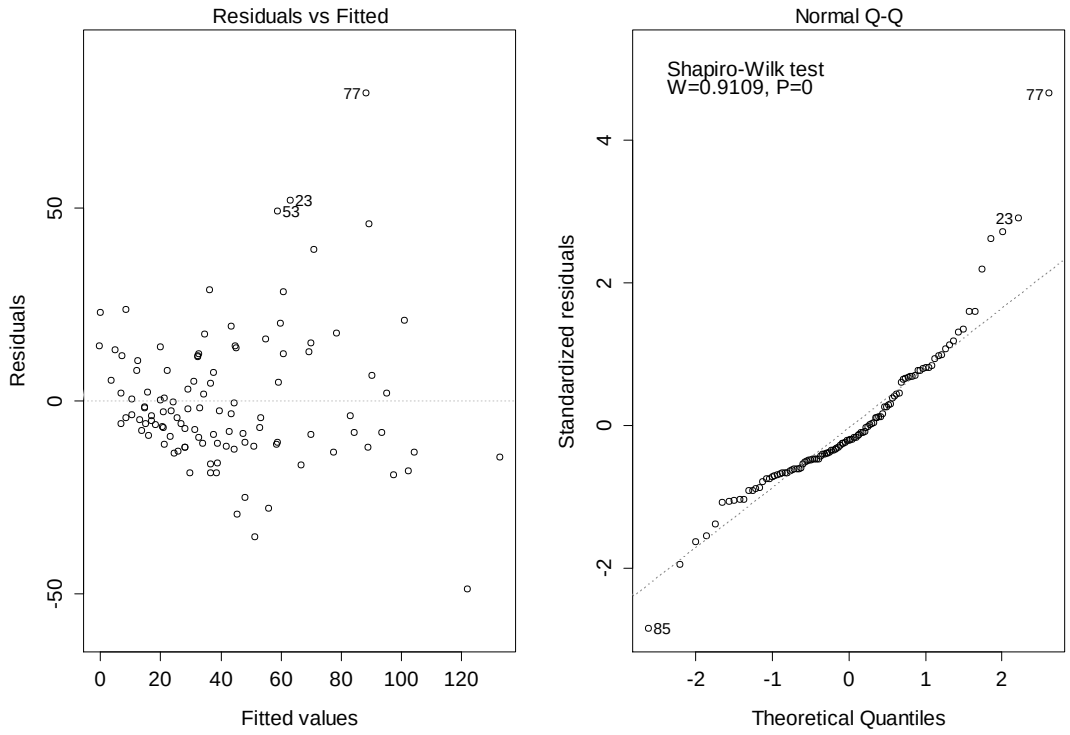
```
wind        -1.337e+01  2.300e+00  -5.810 6.89e-08
I(wind^2)    4.652e-01  1.008e-01   4.617 1.12e-05

Residual standard error: 18.28 on 104 degrees of freedom
Multiple R-squared:  0.7148,       Adjusted R-squared:  0.6984
F-statistic: 43.44 on 6 and 104 DF,  p-value: < 2.2e-16
```

מדוע לדעתכם בחרו החוקרים האת המודל הנ"ל? מה ניתן ללמוד מהפלט? מה טיב התאמת המודל? האם יש צורך **בכל** המשתנים המסבירים? במידה ולא, איזה ניתן לדעתך להוציא ראשון מהמודל?

בגלל הקשר הפרבולי של חלק מהמשתנים המסבירים אל מול המוסבר החוקרים בחרו להתאים מודל של רגרסיה פולינומיאלית. טיב התאמת המודל הוא 0.7148, והמדד המתוקנן הוא בעל ערך 0.6984.

מהפלט אפשר ללמוד שהמודל מובהק, כלומר לפחות אחד המסבירים מסביר את המוסבר. משתנה הרוח מובהק מאד, כמו גם משתנה הרוח (אם כי רמת המובהקות נמוכה יותר).
נראה שאין צורך במשתנה הריבועי של עוצמת קרינת השמש, וספק גדול אם בכלל יש צורך במשתנה הזה.

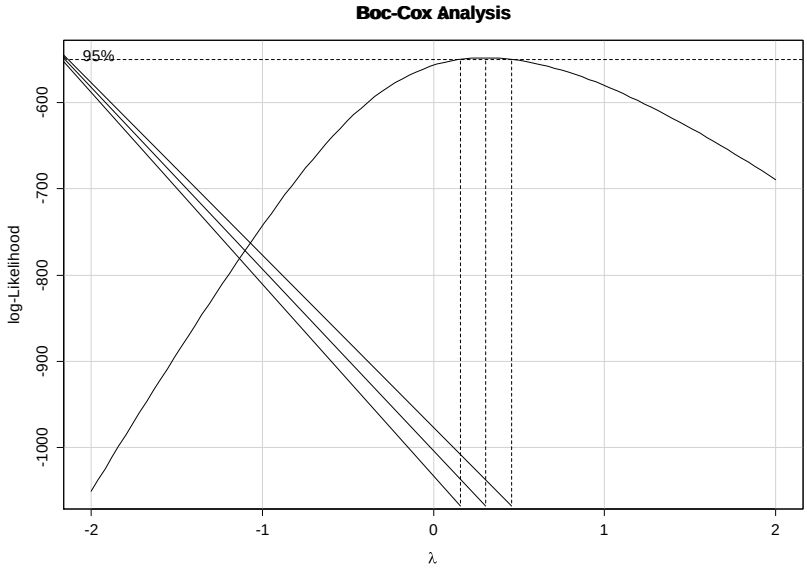
ג. (6 נקודות) איזו שתי הנחות הגרפים הבאים באים לבדוק ומה מסקנתך לגבי הנחות אילו?



לפי תרשים השאריות (השמאלי) נראה שלא מתקיים שיוויון השונויות, בגלל צורת המשפך המתקבל ככל שעולים בערך החזוי. נראה שהמגמה לינארית (כי לא ניתן לזהות מגמה בשאריות) אך ככל שעולים בערכים החזויים, שונותם עולה.

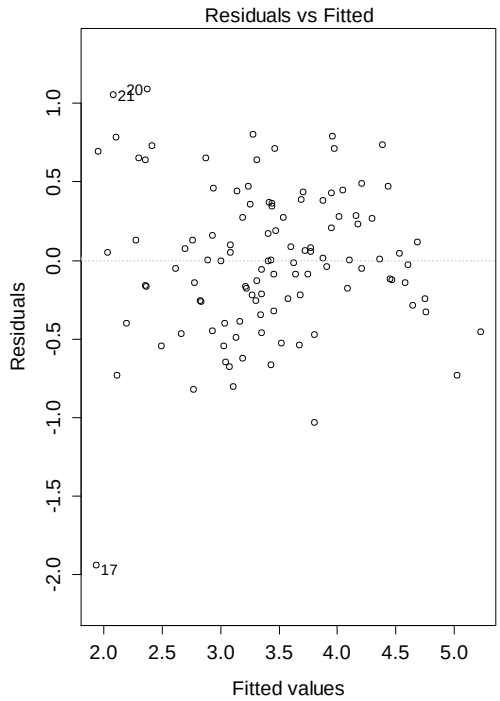
לפי תרשים הכמותנים (הימני) נראה שהנחת הנורמלית אינה מתקיימת כי הנקודות אינן מסתדרות לפי הקו הלינארי, אך עם זאת, לפי מבחן שפירו ווילקס, מתקבל שהנחת הנורמליות כן מתקיימת.

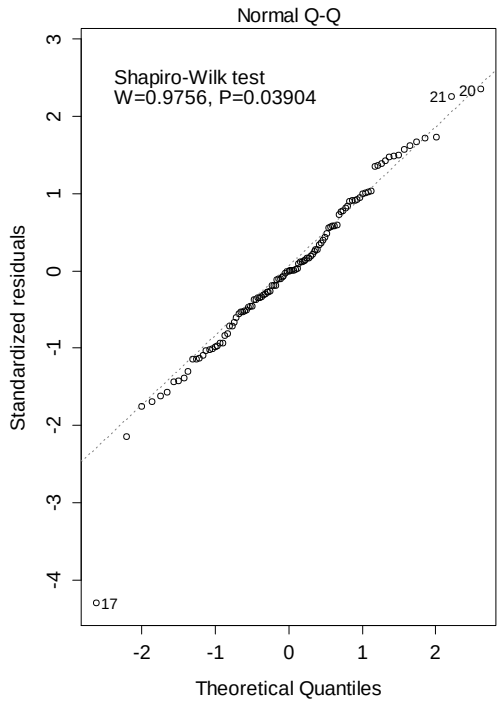
ד. (7 נקודות) החוקרים ערכו את הניתוח שלהלן. מה מטרת הניתוח? מה תהיה המלצתכם לחוקרים?



זהו תרשים בוקס קוקס ומטרתו היא למצוא λ עבור טרנספורמציית החזקה בוקס קוקס, אשר ממקסמת את פונקציית הנראות. מטרת הניתוח היא לשפר את הנחות המודל על ידי טרנספורמצית חזקה (ולוג) על המשתנה המוסבר. נראה שערך הλ הממקסם נמצא בטווח יחסית קטן, בערך באזור ה-0.3. המלצתי לחוקרים הייתה לנסות גם את הערך 0 עבור הטרנספורמציה, היות והוא נמצא בקרבת רווח הסמך (בר"ב נמוכה יותר), על מנת לאפשר פרשנות נוחה של תוצאות המודל (מי שלא כתב את ההמלצה קיבל את מלוא הנקודות).

ה. (7 נקודות) למרות או בגלל התוצאות הקודמות החוקרים החליטו לחזור על הניתוח ולהשתמש ב-log(ozone) כמשתנה תלוי ובדקו שוב את הנחות המודל. מה ניתן ללמוד מתרשימים עכשיו? מדוע חוקר החליט להשמיט את תצפית מספר 17?





לאחר הטרנספורמציה, ניתן לראות שהנחת שיוויון השונויות והנחת הלינאריות מתקיימת, לפי תרשים השאריות, וגם הנקודות בתרשים הכמותונים מסתדרות בצורה טובה יותר על הקו הלינארי, והנחת הנורמליות מתקיימת לא לפי ר"מ במבחן

לאחר הטרנספורמציה, תצפית מס' 17 הוסרה, היות וההתנהגות שלה היא חריגה, והיא מגדילה את השונות עבור ערכים חזויים נמוכים וכפי הנראה פגעה גם בתוצאות מבחן הנורמליות שנעשה ואולי אף השפיעה על האנמ ל-λ

ו. (7 נקודות) מה החוקרים רצו לבדוק בעזרת הפלט הבא? מה תמליצו להם?

```
Analysis of Variance Table

Model 1: log(ozone) ~ rad + I(rad^2) + temp + I(temp^2) + wind +
I(wind^2)
Model 2: log(ozone) ~ rad + temp + wind
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    103 20.175
2    106 23.069 -3   -2.8937 4.9244 0.003086
```

מבחן F חלקי-

$$\begin{cases} h_0: model\ 1 \\ h_1: model\ 2 \end{cases}$$

המבחן בודק אם המקדמים של המשתנים הריבועיים שווים ל-0 או לא. תוצאת המבחן מובהקת ולכן נמליץ להם להשתמש במודל הריבועי (כעדיף על המודל הלינארי).

שאלה מספר 3 (20 נקודות)

נתון מודל רגרסיה מרובה סטנדרטי שבו התצפיות ממוינות לפי אחד המשתנים המסבירים, j_0, כלומר מתקיים $x_{1j_0} \le x_{2j_0} \le \cdots \le x_{nj_0}$. נסמן ב $\boldsymbol{s}$ את הוקטור של סכומי השאריות המצטברים, כלומר $s_k = \sum_{i=1}^{k} e_i$. הסבירו כיצד ניתן להציג נוסחה מטריציונית לחישוב הוקטור $\boldsymbol{s}$. הציגו נוסחאות מטריציוניות לחישוב $E(\boldsymbol{s})$ ו- $COV(\boldsymbol{s})$, תוך שימוש במרכיבים (ערכים/וקטורים/מטריצות) שמוגדרים במודל הרגרסיה המרובה הסטנדרטי. עבור כל מרכיב שבו השתמשתם בנוסחאות, ציינו האם המרכיב ידוע או לא ידוע, וכן האם הוא מקרי או קבוע.

נגדיר מטריצת עזר A, בגודל n על n-

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & & 0 & 0 & 0 & 0 \\ & \vdots & & & \ddots & & \vdots & & \\ 1 & 1 & 1 & 1 & & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$S = A \cdot \underline{e}$$

$$E(S) = E(A \cdot \underline{e}) = A \cdot E(\underline{e}) = \underline{0}$$

$$COV(S) = COV(A \cdot \underline{e}) = A \cdot COV(\underline{e}) \cdot A^t = A\sigma^2(I - H)A^t$$

כל הגדלים למעט ווקטור השאריות הינם קבועים. כולם ידועים למעט σ^2.

בהצלחה!