תרגיל בית 3- קשרים בין משתנים רציפים

מעבר לעזרה שבתוכנה עצמה, מומלץ להיעזר באתר: http://wiki.r-project.org/rwiki/doku.php

הדרכה בעברית ניתן למצוא באתר הקורס.

ניתן ורצוי להתשמש בפורום הקורס להתייעצות.

יש לצרף את הקוד ששימש לפתרון אך אין הוא תחליף לתשובה סופית.

סעיפים בדרגת קושי גבוהה יותר סומנו בכוכבית(*). כדאי להשתדל לענות עליהם אבל אין הכרח.

התאמת קשר לינארי

1. בקובץ sprinttimes.csv שבאתר הקורס נתונים זמני שיא בריצה לנשים ולגברים באולימפיאדות.
   i. שרטטו תרשים פיזור לנתונים כאשר בציר x הוא השנה. מה אתם לומדים מתרשים זה? האם קשר לינארי בין השנה לזמני הריצה הוא סביר לאור התרשים?
   מומלץ להיעזר במתגים pch או col של הפקודה plot בשביל להבחין בין גברים לנשים.
   ii. התאימו קוים חסינים באמצעות הפקודה מתרגול #4 לגברים ולנשים בנפרד. מה השיפועים של הקוים? מה המשמעות (הנורמטיבית, לא המתימטית) של השיפועים?
   iii. התאימו את קווי ריבועים פחותים לנתונים. מה השיפועים שלהם? באיזו שיטת התאמה השיפור של הנשים מידי שנה הוא גדול יותר?
   iv. קראו את המאמר המשווה זמני שיא בריצה בין נשים לגברים שהתפרסם ב-nature ונמצא באתר הקורס. האם אתם מסכימים עם מסקנות הכותבים?
   v. מהי השנה בה הנשים יעקפו את הגברים לפי הקו החסין ולפי קו הריבועים הפחותים?

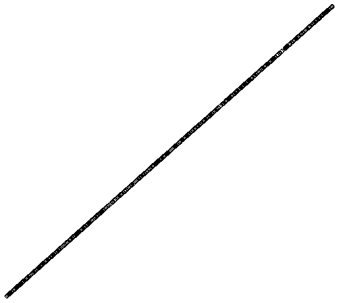
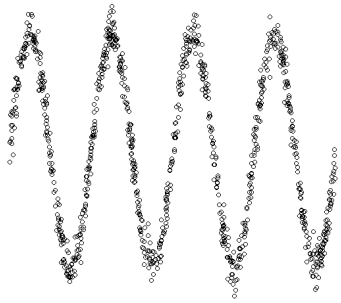
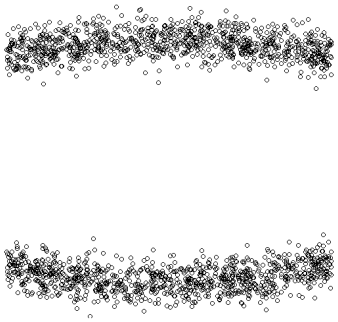
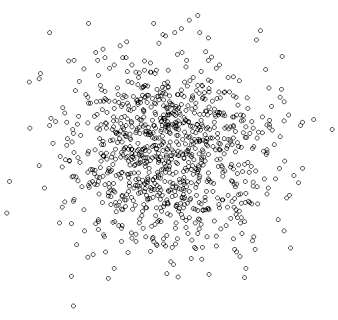
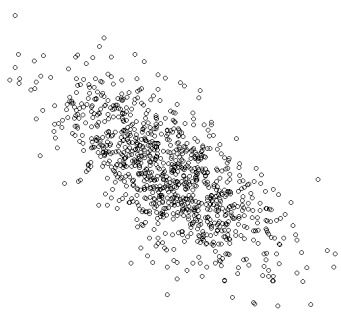
2. בקובץ הנתונים ex03data שבאתר הקורס, ישנם שלושה משתנים: y,x1,x2.
   נניח לצורך הדיון[1] שאלו מחירי חולצות ממותג כלשהו (y) מול מדד זמן כלשהו $(x1)$ ואחוז המודעות למותג בקרב האוכלוסייה $(x2)$.
   (a)התאימו מישור ריבועים פחותים (a.k.a. multiple regression) על ידי הפונקציה lm.
   (b)מהי משוואת הניבוי?
   (c)מסתבר שהקשר האמיתי כלל אינו לינארי... התבוננו בשאריות והציעו טרנספורמציות על המשתנים המסבירים $(x1, x2)$ כך שמישור הניבוי הלינארי יתאים טוב יותר.
   כלומר, אם הקשר האמיתי הוא מהצורה $y=\beta_0+\beta_1\cdot f_1(x1)+\beta_2\cdot f_2(x2)+noise$ עליכם למצוא את f_1, f_2 .
   רמז: f_1, f_2 הן פונקציות אלמנטריות, אך אינן מונוטוניות...
   (d)מה הייתה השונות המוסברת (R^2) לפני הטרנספורמציות ומה השונות המוסברת

1 ויסלחו לי תלמידי הכלכלה על ההפשטה הפלילית של בעיית התמחור.

לאחריהן? האם הטרנספורמציות עזרו להסביר את המחירים?
הערה: מקדם המתאם מוגדר רק עבור מסביר בודד לכן במקרה זה לא ניתן לחשב את השונות המוסברת דרך מקדם המתאם אלא יש לחשב לפי ההגדרה.

3. ועוד על מקדם המתאם...

i. מה תוכלו להגיד על מקדם המתאם (של פירסון) במקרים הבאים? האם חיובי/שלילי? האם קרוב ל ± 1 או לאפס?

A	B	C
		
E	F	G
		

ii. מה תוכלו לאמר על אחוז הפיזור של הנתונים המוסבר על ידי הקשר הלינארי בכל אחד מאותם המקרים?

iii. מה ייקרה למקדם המתאם אם תפעילו את הטרנספורמציות הבאות על הנתונים? אם השתמשתם בהנחות כלשהן בתשובתכם, פרטו אותן.

טרנספורמציה	קובץ נתונים

A	$x_{new}=x^2;\ y_{new}=y$
B	$x_{new}=f(x);\ y_{new}=f(y)$ עבור $f(t)=c+d\cdot t$ כאשר c,d קבועים כלשהם.
G	$x_{new}=x^2;\ y_{new}=y$

התפלגויות רציפות

1. ההתפלגות המעריכית (exponential) נתונה על ידי הצפיפות $f_X(x;\lambda)=\begin{cases}\lambda e^{-\lambda x}, x\geq 0\\ 0, x<0\end{cases}$ ולה פרמטר אי שלילי λ .

   (e)מהי פונקציית ההסתברות המצטברת המעריכית? כלומר מהי הפונקציה $F_X(x)\stackrel{\text{def}}{=}P(X\leq x)$?

   (f)מהי התוחלת של ההתפלגות המעריכית?

   (g)מהי השונות של ההתפלגות המעריכית?

   (h)הראו באמצעות משפט בייז שההתפלגות המעריכית מקיימת את השיוויון $P(X\geq x+t|X>x)=P(X\geq t)$.

   (i)(*) התכונה בסעיף הקודם נקראת תכונת "חוסר הזכרון". הציעו נימוק לשם זה ותרחישים בהם סביר/לא סביר שהתכונה תתקיים.

2. אבנר רוצה לצייר ריבוע (מקרי) על הקרקע. משיקולים השמורים עימו, החליט להגריל את אורך הצלע (X) של הריבוע בצורה שוות סיכוי בקטע [0,1]. כלומר $X\sim Unif[0,1]$.

   (a)מהו הסיכוי שאורך הצלע קטן מאלפא כלשהו? $P(X\leq\alpha)$.
   מהי פונקציית הצפיפות של X?

   (b)מהו הסיכוי ששטח הריבוע $(Y\equiv X^2)$ קטן מאלפא כלשהו? $P(Y\leq\alpha)$.
   מהי פונקציית הצפיפות של המשתנה Y?

   (c)מהי תוחלת גודל הריבוע המקרי? $E(Y)$

   (d)מהי שונות גודל הריבוע? $Var(Y)$

   (e)בשביל לרענן חדו"א 1 , החליט אבנר להגדיר משתנה חדש: $Z\equiv-\ln(X)$.
   מהו הסיכוי ש Z קטן ממספר אלפא כלשהו? $P(Z\leq\alpha)$.
   מהי פונקציית הצפיפות של Z?

   (f)מסתבר ש Z שייך למשפחת התפלגויות מוכרת... לאיזו? מהי שונותו ותוחלתו?