<u>שאלות לדוגמה למבחן:</u>

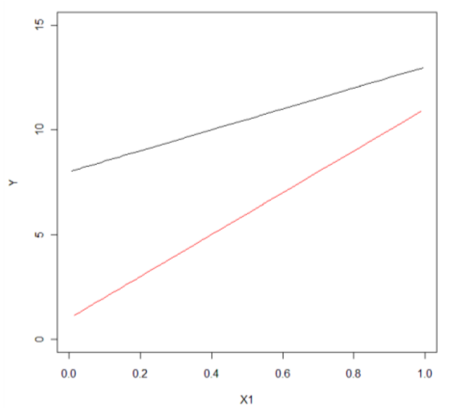
<u>רגרסיה לינארית</u>

1. נתון מודל רגרסיה פשוטה ללא חותך, $i = 1, \dots, n \quad Y_i = \beta_1 X_i + \varepsilon_i$.
   a. חשבו את אומד הריבועים הפחותים: $\hat{\beta}_1$
   b. חשבו את התוחלת והשונות של האומד שמצאתם בסעיף א': $E(\hat{\beta}_1), Var(\hat{\beta}_1)$
   c. הראו כי $2n(MSE_{tr} - MSE_{te}) = 2\sigma^2$. שימו לב שזה מקרה פרטי של משפט שהוכח בכיתה כאשר למודל יש ממד 1. בתשובתכם אין להשתמש ישירות במשפט אבל אפשר (וגם כדאי) להשתמש בכך ש $2n(MSE_{tr} - MSE_{te}) = 2\sum_{i=1}^{n} Cov(Y_i, \hat{Y}_i)$
2. מתאימים מספר מודלים של רגרסיה לינארית להסברת המשתנה Y באמצעות המשתנים המסבירים הבאים: X_1 - משתנה רציף, X_2 - משתנה בינארי

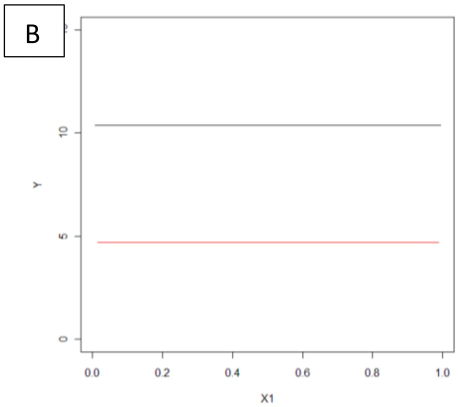
להלן ארבעה מודלים אפשריים וארבעה גרפים המציגים את ערך Y **כתלות בערך X_1**. רשום עבור כל גרף את המודל שהוא מתאר <u>ונמק את תשובתך</u>.

1. $Y = \beta_0 + \beta_1 X_1$
2. $Y = \beta_0 + \beta_2 X_2$
3. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
4. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

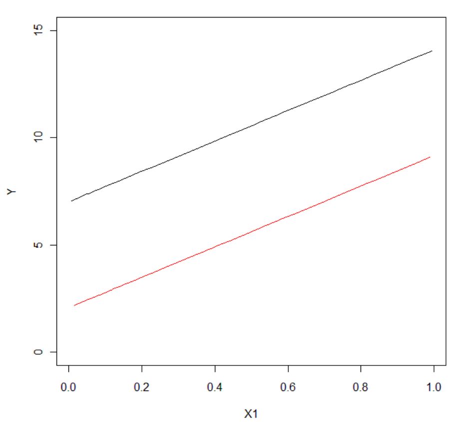
A



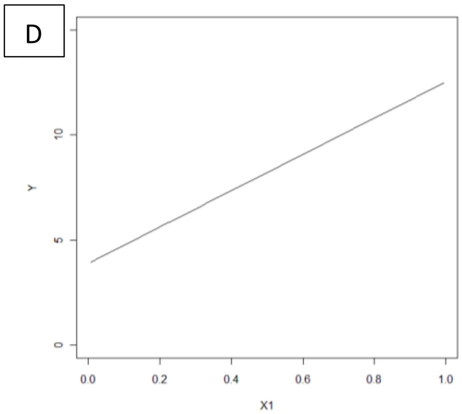
B



C



D



Model selection

1. עבור מודל רגרסיה אנחנו שוקלים שני תת-מודלים: מודל A_p עם p פרמטרים ומודל A_q עם q פרמטרים. נניח שאנחנו רוצים לבחור את המודל שממקסם את ההסתברות האפוסטריורית (*posterior*) שהוא המודל הנכון (הקריטריון של *BIC*). נסמן בπ_p וב-π_q את ההסתברות האפריורית (*prior*) שמודל A_p ומודל A_q הם נכונים בהתאמה.
   a. נניח ש $P(X,Y|A_q) = 2P(X,Y|A_p)$. עבור אילו ערכים של π_p נבחר את מודל A_p ועבור אילו ערכים נבחר את מודל A_q?
   b. האם התשובה תלויה בערכים של *p, q*? הסבר את תשובתך.
   c. עבור נתונים ובהם חמישה משתנים מסבירים $X_1, \dots, X_5$. לאחר סיום תהליך בחירת מודל ע"י ביצוע רגרסיה בצעדים (קדימה ואחורה) בשימוש במדד ההשוואה AIC נבחר המודל הבא: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_5 X_5$. מיהם המודלים שעבורם ערך מדד AIC בהכרח פחות טוב מהמודל הנבחר?

סיווג

1. נתון המודל *Weird-SVM* אשר מתקבל מפתרון בעיית האופטימיזציה הבאה:

$$\min_{(w,b,\xi)} \left(\|w\|^2 + C \sum_{i=1}^{m} I[\xi_i > 0] \right)$$

$$s.t. \quad y_i(\langle w, x_i \rangle + b) \ge 1 - \xi_i$$
$$and\ \xi_i \ge 0\ \forall i$$

   a. הסבר את בעיית האופטימיזציה. השווה אותה לבעיה של *Soft-SVM*.
   b. הנח כי בידיך נתונים ממודל הניתן להפרדה לינארית אך רועש. תאר קובץ נתונים כנ"ל אשר יתקבלו עבורו שני מסווגים שונים מהותית בשני המודלים השונים.

<u>פתרונות:</u>

<u>רגרסיה לינארית</u>

1. נתון מודל רגרסיה פשוטה ללא חותך, $Y_i = \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$.

   a. צריך למצוא β_1 שממזער את $\sum_{i=1}^{n}(Y_i - \beta_1 X_i)^2$. נגזור ונשווה ל-0.

$$\frac{\partial RSS}{\partial \beta_1} - 2\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_i)X_i = 0 \Rightarrow \sum_{i=1}^{n} Y_i X_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

הנגזרת השנייה היא $2\sum_{i=1}^{n} X_i^2 > 0$ ולכן זהו מינימום.

   b. מתקיים ש

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} X_i(\beta_1 X_i + \varepsilon_i)}{\sum_{i=1}^{n} X_i^2} = \frac{\beta_1 \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} X_i\,\varepsilon_i}{\sum_{i=1}^{n} X_i^2} = \beta_1 + \frac{\sum_{i=1}^{n} X_i \varepsilon_i}{\sum_{i=1}^{n} X_i^2} \qquad (*)$$

כיוון שהתוחלת של ε_i היא 0 ו-ε_i ב"ת ב-X_i אז

$$E\left[\frac{\sum_{i=1}^{n} X_i \varepsilon_i}{\sum_{i=1}^{n} X_i^2}\right] = \frac{\sum_{i=1}^{n} E(X_i \varepsilon_i)}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} X_i E(\varepsilon_i)}{\sum_{i=1}^{n} X_i^2} = 0$$

ולכן $E\hat{\beta}_1 = E\left(\beta_1 + \frac{\sum_{i=1}^{n} X_i\varepsilon_i}{\sum_{i=1}^{n} X_i^2}\right) = E(\beta_1) + E\left[\frac{\sum_{i=1}^{n} X_i\varepsilon_i}{\sum_{i=1}^{n} X_i^2}\right] = E(\beta_1) = \beta_1$.

לגבי השונות, כיוון שהביטוי הראשון ב-(*) הוא קבוע אז

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum_{i=1}^{n} X_i \varepsilon_i}{\sum_{i=1}^{n} X_i^2}\right) = \frac{Var(\sum_{i=1}^{n} X_i \varepsilon_i)}{\left(\sum_{i=1}^{n} X_i^2\right)^2} = \frac{\sum_{i=1}^{n} Var(X_i \varepsilon_i)}{\left(\sum_{i=1}^{n} X_i^2\right)^2} = \frac{\sum_{i=1}^{n} X_i^2\, Var(\varepsilon_i)}{\left(\sum_{i=1}^{n} X_i^2\right)^2}$$

$$= \frac{\sum_{i=1}^{n} X_i^2\, \sigma^2}{\left(\sum_{i=1}^{n} X_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$$

   c. לפי הרמז: $2n(MSE_{tr} - MSE_{te}) = 2\sum_{i=1}^{n} Cov\,(Y_i, \hat{Y}_i)$. במקרה שלנו מתקיים ש $\hat{Y}_i =$ $\hat{\beta}_1 X_i = \frac{\sum_{j=1}^{n} X_j Y_j}{\sum_{j=1}^{n} X_j^2} X_i$. ולכן

$$2Cov(Y_i, \hat{Y}_i) = 2Cov\left(Y_i, \frac{\sum_{j=1}^{n} X_j Y_j}{\sum_{j=1}^{n} X_j^2} X_i\right) = \frac{2}{\sum_{j=1}^{n} X_j^2} Cov\left(Y_i, X_i \sum_{j=1}^{n} X_j Y_j\right)$$

$$= \frac{2X_i}{\sum_{j=1}^{n} X_j^2} \sum_{j=1}^{n} X_j Cov(Y_i, Y_j) = \frac{2X_i}{\sum_{j=1}^{n} X_j^2} X_i Cov(Y_i, Y_i) = \frac{2X_i^2}{\sum_{j=1}^{n} X_j^2} \sigma^2$$

כאשר השוויון האחרון נובע מכך ש $Cov(Y_i, Y_j) = \sigma^2$ אמ"ם $i = j$.

לכן

$$2n(MSE_{tr} - MSE_{te}) = \sum_{i=1}^{n} 2Cov(Y_i, \hat{Y}_i) = \sum_{i=1}^{n} \frac{2X_i^2}{\sum_{j=1}^{n} X_j^2} \sigma^2 = \frac{2\sigma^2 \sum_{j=1}^{n} X_j^2}{\sum_{j=1}^{n} X_j^2} = 2\sigma^2$$

2. קו יחיד בעל שיפוע מתאים למוד עם משתנה רציף יחיד. שני קוים אופקיים מקבילים מתאימים למודל עם בינארי יחיד. שני קוים בעלי שיפוע מתאימים למודל על שני המשתנים וללא אינטראקציה. שני קוים בעלי שיפוע שונה מתאימים למודל הכולל אינטראקציה.

<u>Model selection</u>

2. עבור מודל רגרסיה אנחנו שוקלים שני תת-מודלים: מודל A_p עם p פרמטרים ומודל A_q עם q פרמטרים כאשר $p > q$. נניח שאנחנו רוצים לבחור את המודל שממקסם את ההסתברות האפוסטריורית (*posterior*) שהוא המודל הנכון (הקריטריון של *BIC*). נסמן בπ_p וב-π_q את ההסתברות האפריורית (*prior*) שמודל A_p ומודל A_q הם נכונים בהתאמה.

a. ההסתברות האפוסטריורית שמודל A_p הוא הנכון לפי כלל בייס היא

$$P(A_p|X,Y) = \frac{P(X,Y|A_p)\pi_p}{P(X,Y|A_p)\pi_p + P(X,Y|A_q)\pi_q}$$

בדומה, ההסתברות האפוסטריורית שמודל A_q הוא הנכון לפי כלל בייס היא

$$P(A_q|X,Y) = \frac{P(X,Y|A_q)\pi_q}{P(X,Y|A_p)\pi_p + P(X,Y|A_q)\pi_q}$$

לכן, נבחר במודל A_q אם ורק אם $P(X,Y|A_q)\pi_q > P(X,Y|A_p)\pi_p$.

נציב $\pi_q = 1 - \pi_p$ וכן $P(X,Y|A_q) = 2P(X,Y|A_p)$ ונקבל שנבחר במודל A_q אם ורק אם

$$2P(X,Y|A_p)(1-\pi_p) > P(X,Y|A_p)\pi_p \Rightarrow 2 - 2\pi_p > \pi_p \Rightarrow \frac{2}{3} > \pi_p$$

b. כפי שראינו בפיתוח, התוצאה אינה תלויה בערכי הפרמטרים p,q. בפיתוח שראינו בהרצאה מספר הפרמטרים הגיע כחלק מקירוב של ההסתברות $P(X,Y|A_k)$.

c. כאשר מבצעים רגרסיה בצעדים לא מתכנסים בהכרח לפתרון אופטימלי מכיוון שהאלגוריתם חמדני ולכן לא ניתן לומר כי מדד AIC של המודל הנבחר טוב משל כל המודלים האחרים. עם זאת כל המודלים אשר נבדלים מהמודל הנבחר במשתנה מסביר יחיד בהכרח פחות טובים מהמודל הנבחר כי אחרת היה מתבצע צעד נוסף.

$Y = \beta_0 + \beta_3X_3 + \beta_5X_5$

$Y = \beta_0 + \beta_1X_1 + \beta_5X_5$

$Y = \beta_0 + \beta_1X_1 + \beta_3X_3$

$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_5X_5$

$Y = \beta_0 + \beta_1X_1 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5$

<u>סיווג</u>

2. נתון המודל *Weird-SVM* אשר מתקבל מפתרון בעיית האופטימיזציה הבאה:

$$\min_{(w,b,\xi)}\left(\|w\|^2 + C\sum_{i=1}^{m} I[\xi_i > 0]\right)$$

$$s.t.\quad y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$$
$$and\ \xi_i \geq 0\ \forall i$$

a. הניסוח של בעיית האופטימיזציה הנ"ל הינו למצוא מינימום לפונקציה ובה שני איברים ובניהם קיים *trade-off* אשר ניתן לכיוון באמצעות הפרמטר *C*. האיבר השמאלי בפונקציית המטרה מבטא חיפוש של מפריד בעל מרווח (*margin*) מקסימלי בין שתי המחלקות ואילו האיבר השני מבטא חיפוש של מפריד בעל **מספר שגיאות** מינימלי. ההבדל בין הבעיה הזאת לבעיית *Soft-SVM* הינו בכך שבזאת מחפשים למזער את מספר השגיאות בעוד שב*Soft-SVM* ממזערים את סכום השגיאות.

b. מעט תצפיות אשר חורגות באופן משמעותי (בערכי המשתנים המסבירים) מההתפלגות של המחלקה שלהן לכיוון המחלקה השנייה.