סדנת Machine Learning

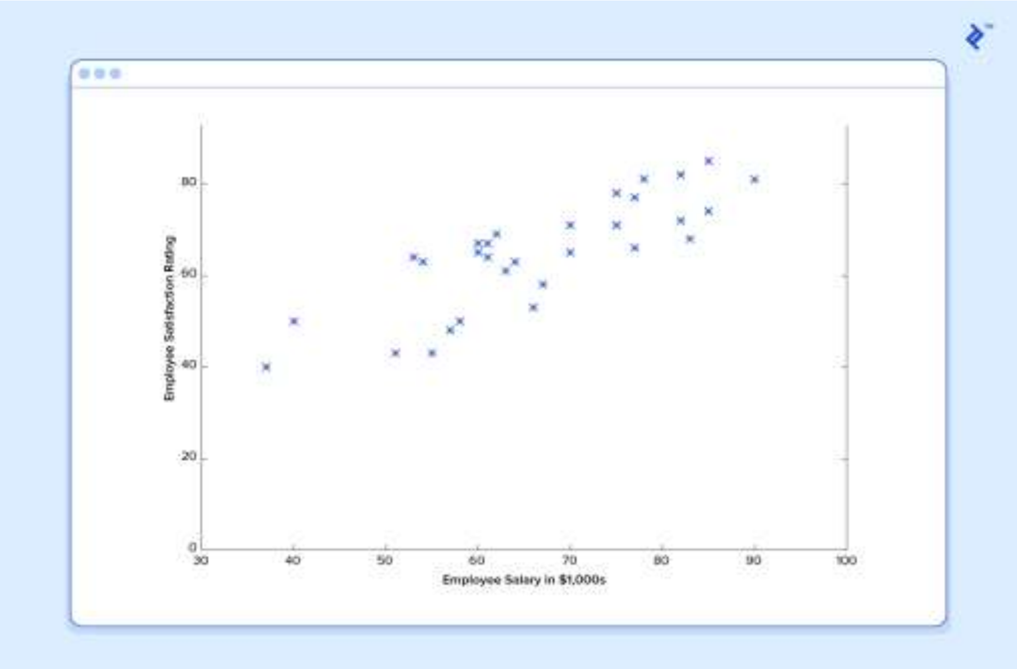
מפגש ראשון- Linear Regression

חלק מהשקפים נלקחו מהספר:
"An Introduction to Statistical Learning"

חזרה

<u>דוגמא</u>:

נניח ובידינו נמצאים הנתונים הבאים על שביעות רצון העובדים (בסקאלה של 1-100) ורמת השכר שלהם, כמוצג בגרף הבא:



אנו מעוניינים <u>לחזות</u> בהינתן רמת השכר את שביעות רצון העובד.

ראשית, ניתן לשים לב שהדאטא מעט רועש ולא אחיד, ועם זאת אנו מסוגלים לראות איזשהו קו מנחה שמראה כי ככל שהשכר עולה רמת שביעות הרצון עולה.

חזרה

דוגמא:

לאחר שהבנו כיצד המידע "מתנהג" עלינו לבנות מודל שיתאר אותו. בהתאם למודל שהצגנו מקודם נייצג את המודל בצורה הבאה : $h(x) = \theta_0 + \theta_1 x$

x- שכר העובד

θ_0- משתנה חופשי (שולט על ה"רמה ההתחלתית של שביעות הרצון –חותך לציר ה-X)

θ_1- מקדם של משתנה השכר (בעצם כמה משתנה השכר משפיע על שביעות הרצון)

$h(x)$ - ערך שביעות הרצון החזוי (ניתן להתייחס אליו כ-$\hat{y}$)

חזרה

דוגמא:

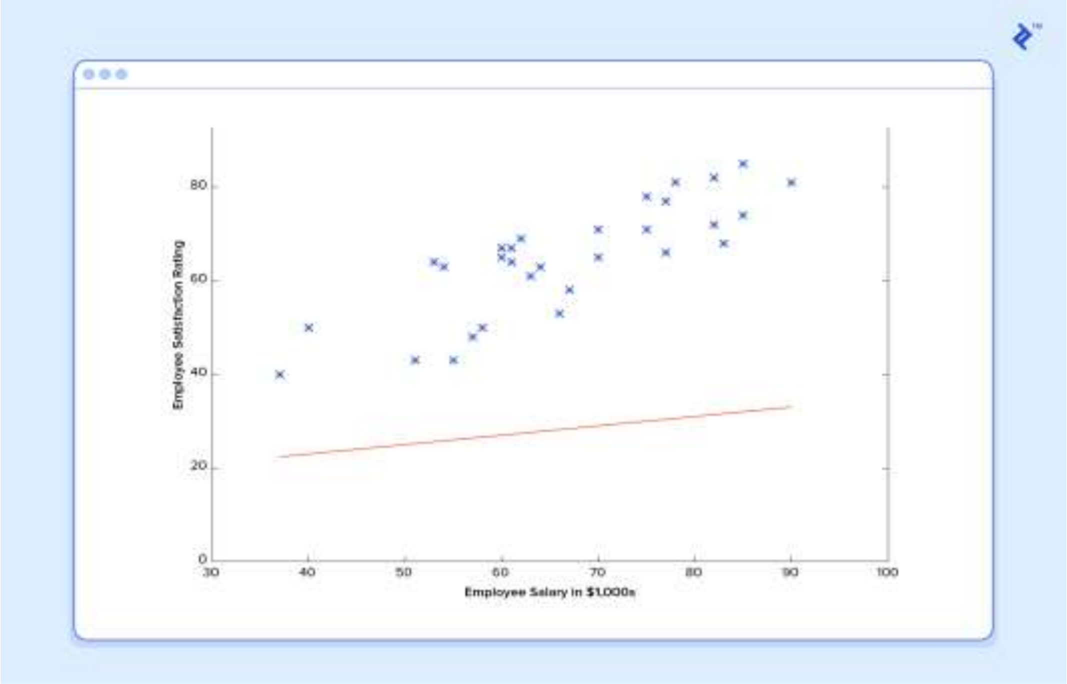
שלב האתחול:

נאתחל את המודל בצורה בסיסית עם ערכים של:

$\theta_1 = 0.2$ ו-$\theta_0 = 12$

ונקבל: $h(x) = 12 + 0.2x$.

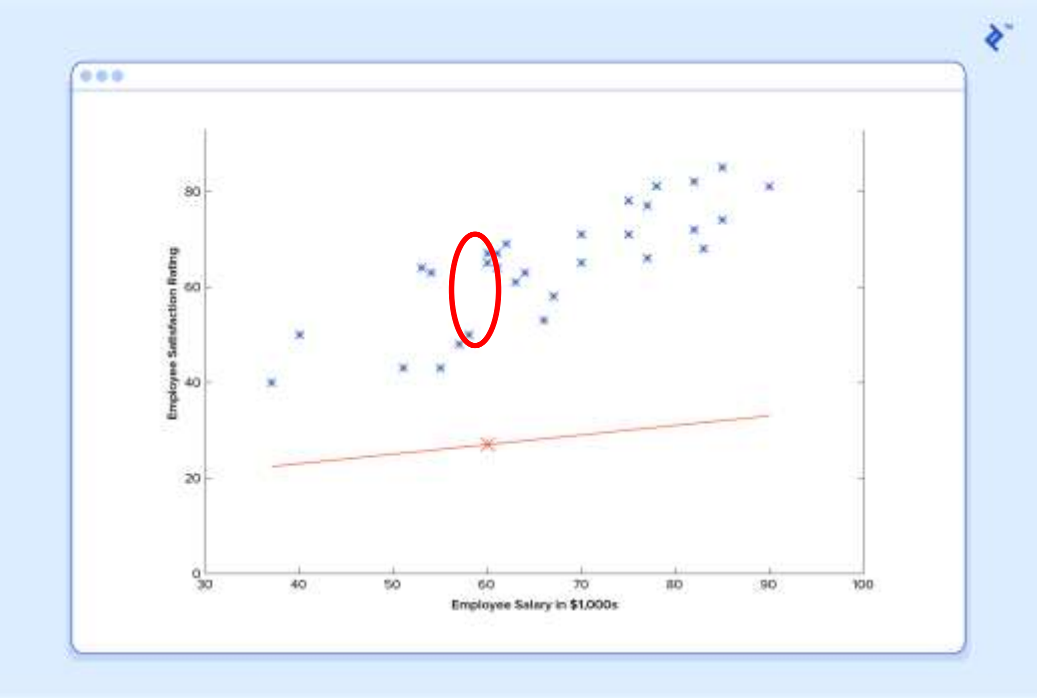
נייצג את המודל שקיבלנו בצורה גרפית ונקבל:



חזרה

<u>דוגמא</u>:

אם נבקש מהמודל לבצע פרדיקציה (תחזית) לרמת שביעות הרצון של עובד ששכרו הוא $60k נקבל את התוצאה הבאה:

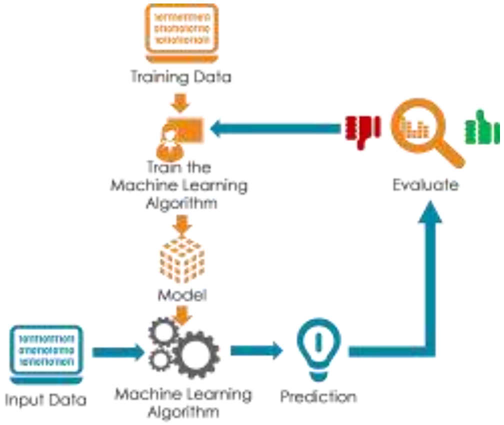


ניתן לומר בבירור שהמודל לא קרוב לתוצאה האמיתית (27 חזוי מול 50-60).

מה עושים כדי לשפר את המודל? <u>נאמן אותו</u>!

דוגמא:

שלב האימון :



נרצה לראות **כמה** אנו רחוקים מהתוצאות האמיתיות **ולתקן** בהתאם. לצורך כך:

1. נעבור על **כל** הדגימות שברשותנו ממדגם האימון.
2. נבצע תחזית עבור כל אחד ונמדוד את ה**שגיאה.**
3. נסכום את כלל השגיאות ונמצא את השגיאה הממוצעת
4. נבדוק (באמצעות גזירה – יוסבר לעומק בהרצאות הבאות) כמה עלינו לשנות כל משתנה במודל כך שימזער את השגיאה
5. נחסיר/נוסיף לכל משתנה בהתאם לתוצאות שקיבלנו
6. נחזור לשלב הראשון

חזרה

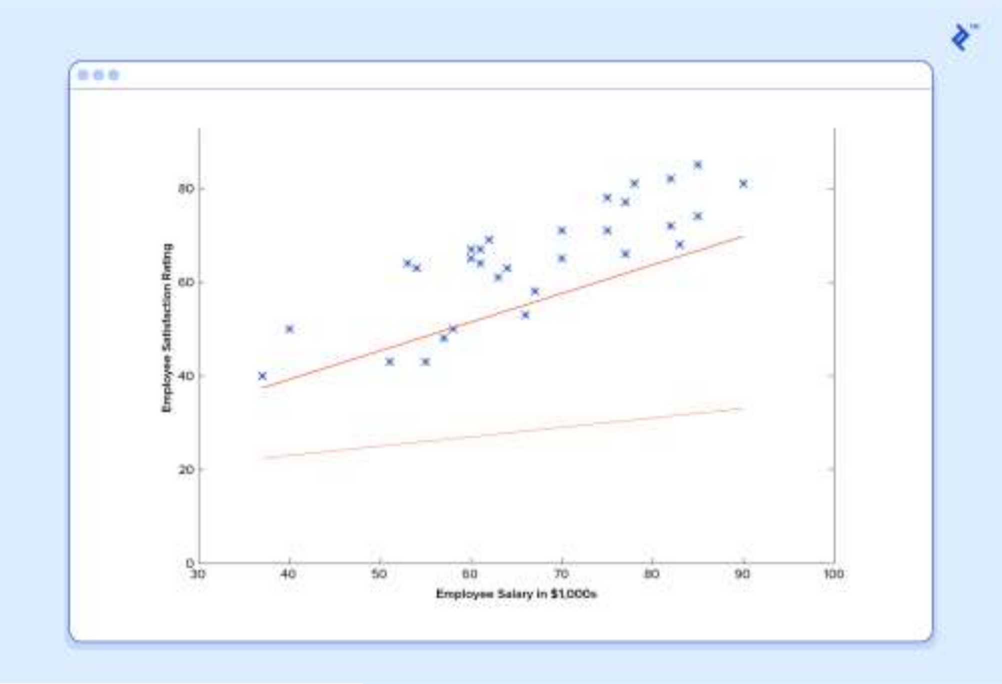
דוגמא:

לאחר שביצענו סבב אימון אחד. קיבלנו שעבור ערכים של $\theta_0 = 13.12$ ו-$\theta_1 = 0.61$, המודל יהיה יותר קרוב לתוצאות האמת. נכתוב את המודל החדש:

$$h(x) = 13.12 + 0.61x$$

נייצג את המודל שקיבלנו בצורה גרפית ונקבל:

ניתן לראות שהתוצאות כבר סבירות.

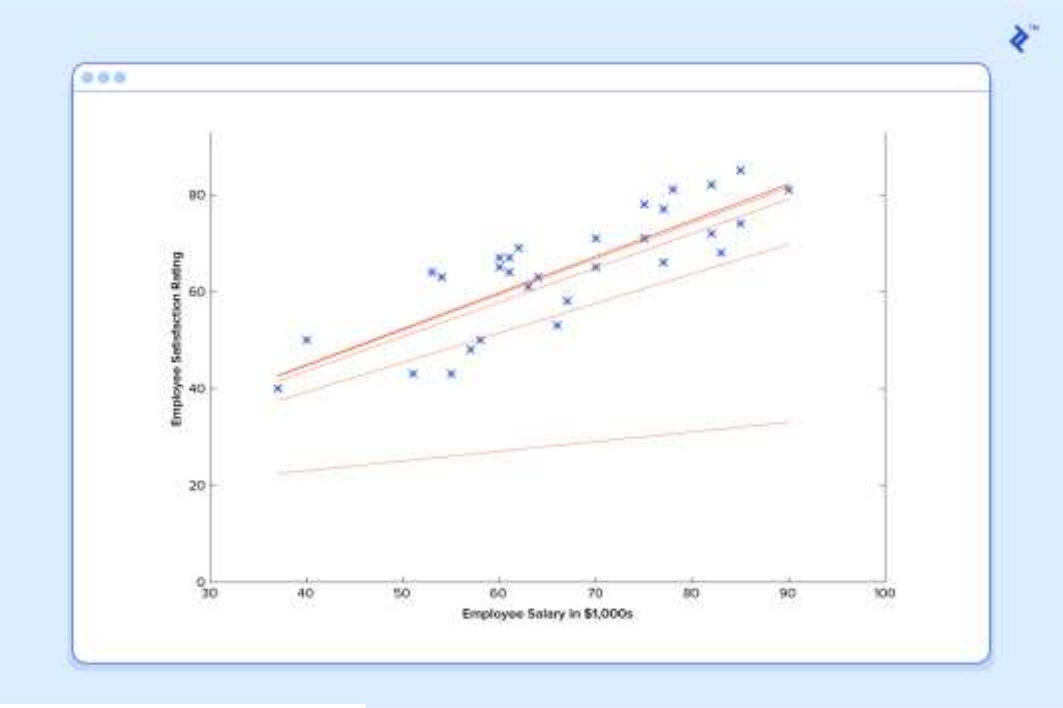


- <u>דוגמא</u>:

אם נחזור על תהליך האימון המון פעמים (נניח 1500 איטרציות) נקבל את המודל הבא:

$$h(x) = 15.54 + 0.75x$$

נייצג את המודל שקיבלנו בצורה גרפית ונקבל:



עכשיו זה כבר נראה מעולה.

מה יקרה אם נחזור עוד פעם על תהליך האימון?

הערכת שגיאת המודל

חלק מהעניין בתחום למידת המכונה הוא היכרות עם המון מודלים שונים. מדוע אנו לא משתמשים במודל אחד שהוא הכי טוב?

No free launch in statistics-לכל מודל ישנם יתרונות וחסרונות כאשר האתגר הוא למצוא את המודל המתאים ביותר למידע שבידינו

כיצד נשווה בין מודלים שונים?

אנו זקוקים לכלי להערכת השגיאה/חוסר ההתאמה של המודל כך שייתן לנו מידע לגבי מידת הנכונות שלו.

גם כאן ישנן מגוון שיטות ועלינו לבחור את השיטה הנכונה עבורנו.

הערכת שגיאת המודל

<u>שגיאה ריבועית ממוצעת</u>

<u>זוהי השיטה הבסיסית ביותר:</u>

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

אנו ממצעים את השגיאה בין הערך החזוי לבין הערך האמיתי. ככל שהערך יותר נמוך השגיאה קטנה יותר והמודל יותר מוצלח.

אנו מחשבים את הערך הזה על מדגם האימון (מה שלא כ"כ מעניין אותנו) ועל מדגם המבחן.

ההנחה היא שככל ששגיאת האימון יורדת כך גם שגיאת המבחן (האם בהכרח?)

<u>ישנה הגדרה דומה גם עבור סיווג:</u>

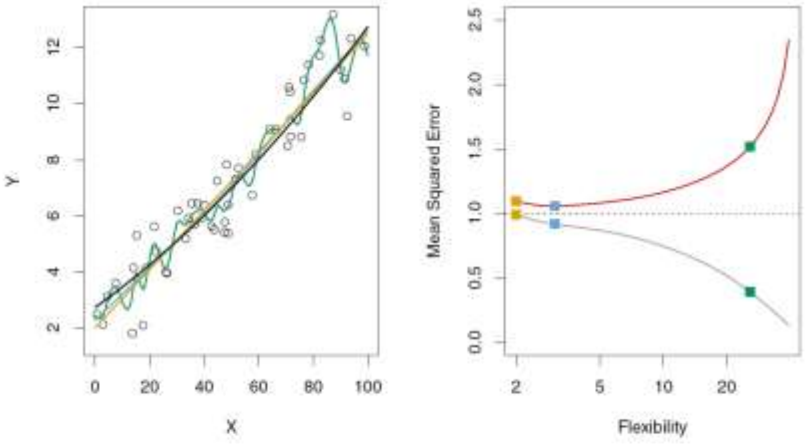
$$\frac{1}{n}\sum_{i=1}^{n}I(y_i \neq \hat{y}_i)$$

אנו סופרים את מספר הסיווגים הלא נכונים שלנו וממצעים אותו.

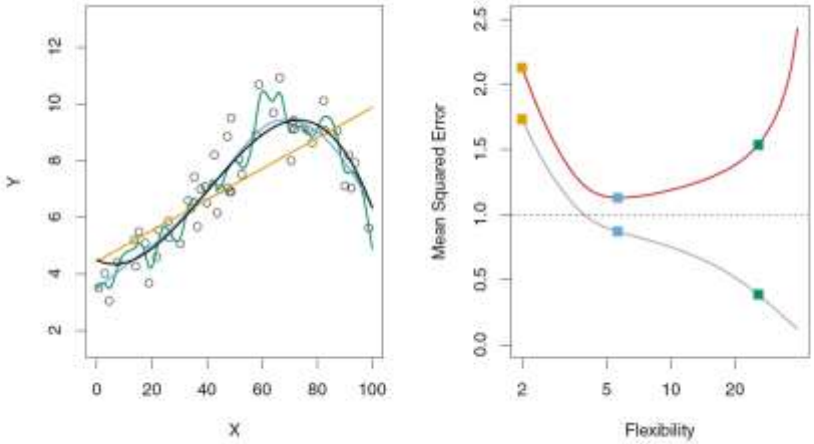
Overfitting

נשים לב לשני הפאנלים. בשניהם ניתן לראות בצד הימני מעין צורת U, הצורה הזו מלמדת אותנו שבשלב מסויים השגיאה על מדגם המבחן עולה למרות ששגיאת האימון יורדת. התופעה הזו נקראת: ***overfitting***
ישנן דרכים להתמודד עם בעיה זו ונדון בהם בהמשך הסדנה

נתונים ליניאריים



נתונים לא-ליניאריים



Bias vs Variance Tradeoff

בהמשך לצורת ה-U שראינו בגרף של שגיאת המבחן, ניתן להבין כי הצורה המיוחדת הזו נובעת משני סטטיסטיים ה"מתחרים" ביניהם: $Bias\&Variance$. ניתן להראות כי ע"י פעולות מתמטיות ניתן לבטא את תוחלת השגיאה כך:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

אם נסתכל על המרכיבים השונים נראה כי הביטויים חיוביים (שונות תמיד חיובית וההטיה בריבוע) מה שאומר ששגיאת המבחן לעולם תקטן עד לשגיאה הבלתי ניתנת להסרה ($irreducible\ error$ זוכרים?)

<u>מהם שונות והטיה?</u>

<u>שונות-</u> כמה **שונים** נתוני האימון בין המדגמים, במילים אחרות, אם אתן למודל נתונים שונים **כמה הם ישפיעו** על הפונקציה הסופית (מה יקרה לדעתכם למודל הירוק מהשקף הקודם?) ככל שהמודל ישתנה יותר כך נאמר שיש לו <u>שונות גבוהה</u>

<u>הטיה-</u> כמה המודל שלנו מתאר בצורה מדוייקת את המודל האמיתי של הנתונים (בד"כ העולם לא לינארי...) ככל שהמודל יתאים בצורה מדוייקת יותר כך נאמר שיש לו <u>הטיה נמוכה</u>

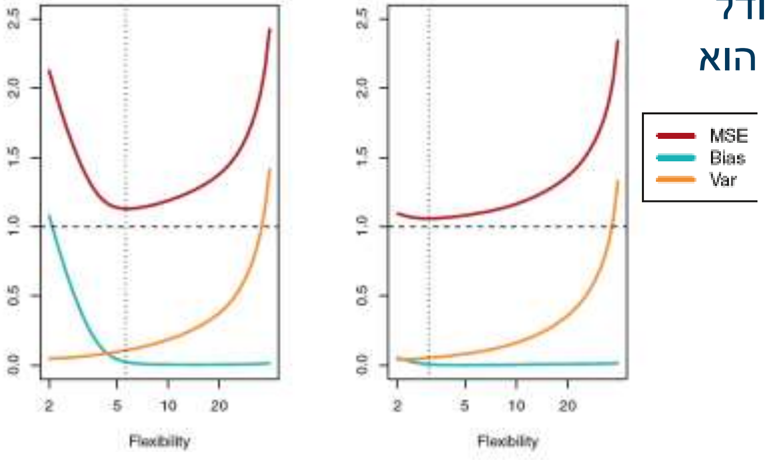
Bias vs Variance Tradeoff

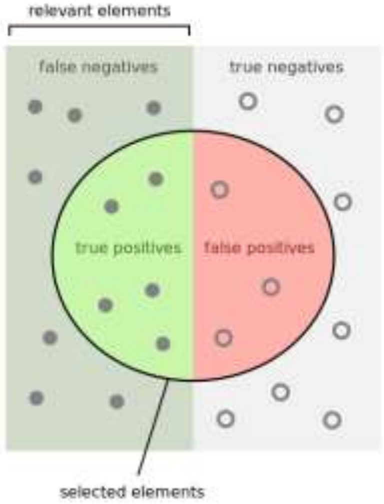
ככלל, אם נשתמש במודלים יותר גמישים נקבל שונות גבוהה יותר והטיה נמוכה יותר. השינוי היחסי בין המרכיבים הללו קובע את גודל שגיאת המבחן.

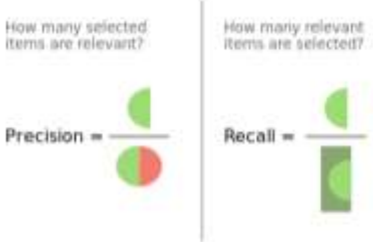
אם נבחר במודל גמיש, ככל שנאמן ההטיה תקטן במהירות והשונות תגדל לאט, עד שנגיע לנקודת המינימום (U) ושם הכיוון יתהפך.

אנו קוראים לתכונה זו tradeoff בגלל שזה יחסית קל למצוא מודל שנותן שונות נמוכה (רעיון?) או הטיה נמוכה (...?) אבל המחיר הוא פגיעה בצד השני.

בתמונה ניתן לראות את הפירוק למרכיבים עבור הגרפים מהשקף הקודם וכיצד השילוב שלהם מייצר את שגיאת המבחן.







<u>נתאר לעצמנו את המקרה הבא:</u>

אנו מעוניינים לחזות תקיפת סייבר על בסיס התראות מסויימות כאשר ידוע ש99% מההתראות הן התראות שווא. אם נחזה בכל פעם שאין תקיפה נקבל מודל שמדייק ב99%. נשמע טו. לא?

ממש לא, אנו נפספס את הנקודה המרכזית שלשמה אנו מבצעים את החיזוי. לכן אנו צריכים <u>מדד אחר</u>.

ישנם שני מדדים מקובלים:

<u>*Recall*</u>- כמה מתוך המקרים שהוגדרו <u>*True*</u> אכן סווגו נכון?

<u>*Presicion*</u>- כמה מתוך מה שסווג כנכון הוא אכן <u>*True*</u> ?

<u>**ישנן כמובן המון שיטות נוספות**</u>

פתיחה

היום נעסוק במודל הנקרא : רגרסיה לינארית.

- זהו מודל מאוד פשוט ללמידה מונחית כאשר פונקציית המטרה היא נומרית (מספרית).
- השיטה קיימת כבר זמן רב (זוהי שיטה סטטיסטית בסיסית) ולעיתים עלולה להיראות מיושנת ביחס לשיטות חדשות ומתקדמות, אך היא עדיין שימושית בהמון מקרים ומייצרת שיטת בסיס שאליה ניתן להתייחס
- שיטות מתקדמות יותר (שלעיתים פשוט מרחיבות את המודל הבסיסי) כך שלימוד שלה בצורה מעמיקה ורחבה יכול להועיל גם כאשר מתמודדים עם מודלים ושיטות מתקדמים יותר

פתיחה

מה נלמד היום?

רגרסיה לינארית פשוטה

- הערכת פרמטרים
- הערכת דיוק הפרמטרים
- הערכת דיוק השגיאה

רגרסיה מרובה

- הערכת פרמטרים

שאלות ודוגמאות

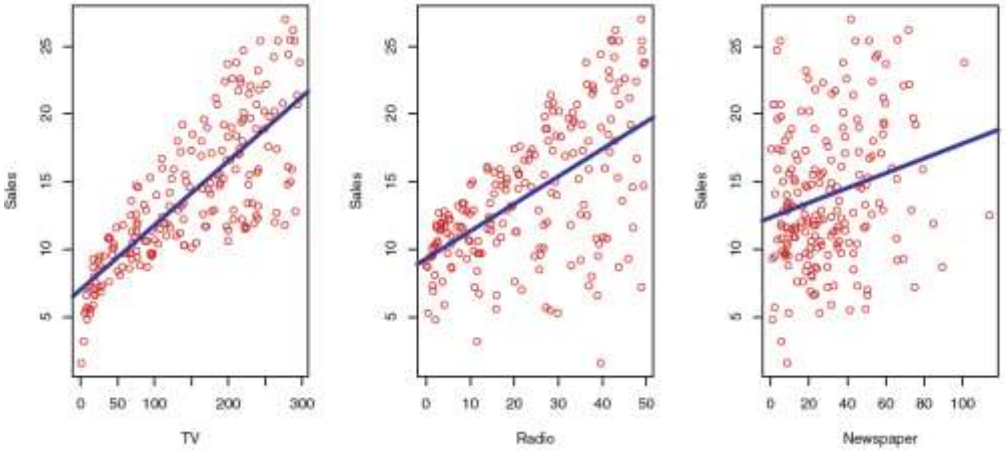
רגרסיה לינארית פשוטה

נתחיל באמצעות דוגמא מלווה, נשאל מספר שאלות וננסה לענות עליהן תוך כדי התבססות על מודל של רגרסיה לינארית.

באמצעות התהליך ננסה ללמוד על מאפיינים חשובים של המודל יתרונות וחסרונות וייצוג נכון של הבעיה.

<u>נסתכל על הדוגמא הבאה:</u>

אלו הם נתוני מכירות של מוצר (במאות אלפים) כפונקציה של אמצעי פרסום שונים (רדיו, טלוויזיה ובעיתונות)



רגרסיה לינארית פשוטה

נניח ונשכרנו (כ-סטטיסטיקאים מומחים) בידי החברה ,זאת במטרה להמליץ על מודל פרסום בשנה הקרובה כך שיניב היקף מכירות גבוה ככל הניתן. איזו אינפורמציה תהיה קריטית לשאלה הזו?

<u>לצורך כך אנו זקוקים לדעת תשובות למספר שאלות:</u>

- האם יש כלל קשר בין פרסום (כלשהו) לבין המכירות ? (אולי לא כדאי בכלל לפרסם?)
- בהנחה שישנו קשר בין תקציב הפרסום לבין המכירות (והתקציב ידוע), האם אנו יכולים לחזות (בדיוק גבוה) את היקף המכירות?
- איזו מדיה תורמת להגדלת היקף המכירות? האם כולן? האם אנו יודעים להעריך זאת ברמת דיוק גבוהה?
- האם הקשר בין הפרסום לבין המכירות הוא לינארי?
- האם ישנו קשר בין אמצעי הפרסום השונים? (למשל: יש לי תקציב של 100K$ כיצד לחלק אותו נכון ?)

רגרסיה לינארית פשוטה

רגרסיה לינארית פשוטה מתייחסת למקרה בו אנו בונים מודל בו אנו חוזים ערך (Y) על בסיס פרמטר **בודד** (X)

המודל מניח שישנו קשר לינארי בין שני המשתנים.

בצורה מתמטית נוכל לכתוב זאת כך:

$$Y \approx \beta_0 + \beta_1 X$$

בדוגמא שלנו אם ניקח את המכירות כערך החזוי (y) ואת הפרמטר להיות טלוויזיה נקבל:

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

β_0 ו - β_1 הם שני משתנים חופשיים לא ידועים, המייצגים את ה<u>חותך</u> וה<u>שיפוע</u> בהתאמה ונקראים **coefficients**

כאשר נשתמש בנתוני האימון שלנו נוכל לקבל הערכה למשתנים שלנו ובהתאם נקבל ערך חזוי:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

הסימון $\hat{}$ מציין אפרוקסימציה (הערכה) למשתנה –נשתמש בו הן לערך החיזוי והן למשתנים

רגרסיה לינארית פשוטה

כאשר נבצע רגרסיה לינארית (בתוכנה כזו או אחרת) נקבל בדרך כלל את הפלט הבא (או דומה לו):

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

מה אנו יכולים להבין מהפלט הזה? מהי המשמעות של המספרים?

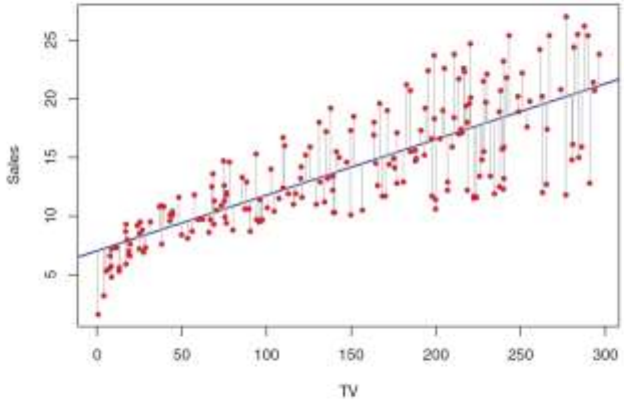
על מנת להבין את הטבלאות הללו אנו נדרש למעט הקדמות...

רגרסיה לינארית פשוטה

הערכת המקדמים

מכיוון שהפרמטרים אינם ידועים אנו נדרשים לבצע הערכה שלהם.

נניח שהדגימות שלנו נראות כך: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, כאשר X מייצג תקציב הפרסום בטלוויזיה ו-y את ערך המכירות ב-200 נק' שונות (n) אנו מעוניינים למצוא את המשתנים β_0 ו - β_1 כך שנקבל את הקו הלינארי בעל השגיאה הנמוכה ביותר תחת שגיאה המוגדרת כ: **Least squares** (האם ישנה הגדרה אחת לשגיאה?)



נגדיר: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ - הערך החזוי עבור דגימה.

$e_i = y_i - \hat{y}_i$ - ההפרש בין הערך החזוי לבין הערך המקורי **(Residual)**

$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$ - סכום ההפרשים בריבוע ניתן לכתיבה גם כ:

Residual Sum of Squares

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

רגרסיה לינארית פשוטה

הערכת המקדמים

שימוש ב- Least squares מנסה למזער את השגיאה באמצעות בחירת β_0 ו - β_1 שייתנו שגיאה מינימלית:

כפי שאנו יודעים, על מנת למזער פונקציה (גזירה) אנו יכולים לגזור לפי כל אחד מהמשתנים ולמצוא את ערך המינימום.

אם נגזור את הערכים ונשווה ל-0 נקבל :

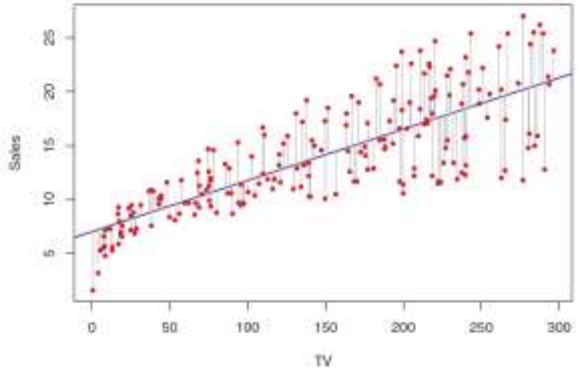
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

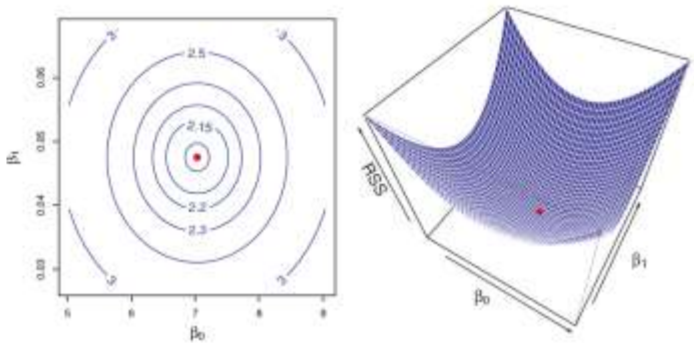
כאשר: $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$

$$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

רגרסיה לינארית פשוטה



$$\hat{\beta}_1 = 0.0475$$
$$\hat{\beta}_0 = 7.03$$



<u>המחשה:</u>

לפי הנתונים של הדוגמא שלנו, מצאנו את <u>הקו הקרוב ביותר</u> לכל הנקודות כאשר:

הנתון של β_1 אומר כי בערך לכל 1000 שנשקיע נמכור עוד 47.5 יחידות.

<u>באיורים למטה</u> ניתן לראות את השגיאה כתלות במשתנים β_0 ו - β_1.

נשים לב כי כפי שתיארנו, הערך המינימלי מתקבל בנקודת המינימום

רגרסיה לינארית פשוטה

<u>הערכת הדיוק של מקדמי המשוואה</u>

ניזכר כי אנו מעוניינים למצוא את היחס בין X ל-Y כאשר אנו מודעים לשני נתונים:

- בידינו מדגם מהמידע ולא את כל המידע
- המידע איננו מתנהג בהכרח בצורה לינארית

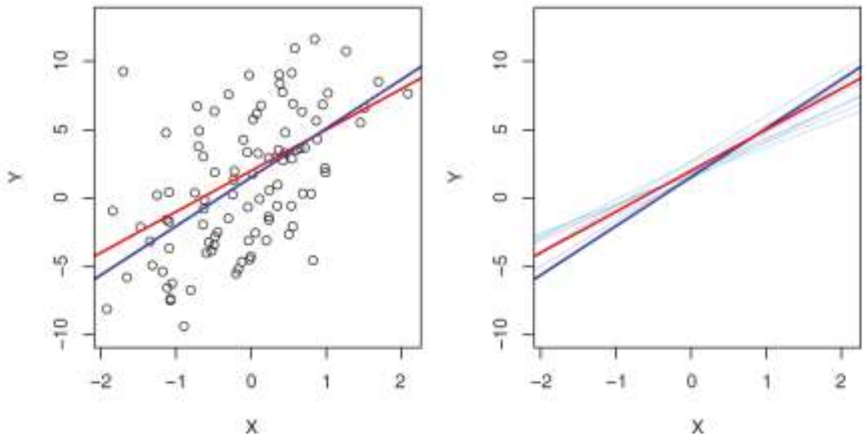
<u>כיצד משפיעות הידיעות האלו על דיוק המקדמים?</u>

אם נביט בתמונות משמאל, נוכל לראות:

<u>קו אדום</u>- התפלגות הנתונים המקורית

<u>קו כחול כהה</u>- קו הרגרסיה על בסיס המדגם שלנו

<u>קו כחול בהיר</u> – קווי רגרסיה על בסיס מדגמים שונים



רגרסיה לינארית פשוטה

הערכת הדיוק של מקדמי המשוואה

אז קצת סטטיסטיקה...

אמד – משתנה ה**אומד** משתנה אחר שאיננו ידוע.

לדוג': אנו מעוניינים לדעת ממוצע של משתנה כלשהו Y, כאשר יש לנו מספר דגימות מההתפלגות של Y. אמד הגיוני לממוצע של Y יהיה ממוצע הדגימות שברשותנו. האם האמד יהיה זהה לממוצע? **לא, אבל..**

ההנחה היא שככל שנקבל יותר דגימות מY נוכל להתקרב לממוצע האמיתי-> זהו אמד חסר הטיה- אין הטיה קבועה למעלה או למטה.

נחזור אלינו:

כמו בממוצע כך גם במקדמי המשוואה, אנו יכולים לבצע הערכה על בסיס המדגם שלנו, ככל שיהיו יותר דגימות כך האמדים למקדמים יהיו יותר מדוייקים.

רגרסיה לינארית פשוטה

<u>הערכת הדיוק של מקדמי המשוואה</u>

<u>אז איך נדע בדיוק **כמה** מדוייקים אנחנו ?</u>

שגיאת תקן, $\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$ זוהי נוסחה סגורה לשגיאת תקן עבור משתנה הממוצע (אמד בדוגמא שלנו)

כאשר σ מייצג <u>סטיית תקן</u> של כל דגימה y_i ו-n כמות הדגימות.

<u>נשים לב</u>, ככל ש -n עולה שגיאת התקן קטנה ולהפך.

שגיאת תקן היא כמות הסטייה הממוצעת של האמד מהמשתנה אותו אנו אומדים

<u>כעת, בצורה דומה נוכל להכליל למקדמי המשוואה:</u>

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

<u>כאשר:</u> $\sigma^2 = \mathrm{Var}(\epsilon)$ תחת הנחה כי ϵ מתפלג עם תוחלת 0 והמשתנים בלתי תלויים

רגרסיה לינארית פשוטה

הערכת הדיוק של מקדמי המשוואה

במקרה הכללי σ כמובן איננו ידוע ולכן אנו זקוקים לבצע הערכה אף אליו:

$$\underline{\textit{residual standard error}}$$
$$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$$

לכן כשאנו מבצעים הערכה ל σ על בסיס המדגם שבידינו נכון לכתוב:
$\widehat{\mathrm{SE}}(\hat{\beta}_1)$ (הערכה לסטיית תקן).

רווח סמך:

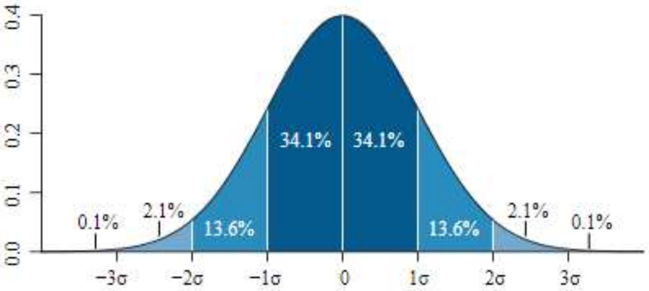
סטיית תקן עוזרת לנו לחשב רווחי סמך- רווח סמך ב95% ביטחון: תחום ערכים כך שב95% תחום זה כולל את הערך האמיתי (הלא ידוע) אותו אנו מחפשים.

התחום הוא משני הצדדים של הערך ומוגדר כך (למקדמי המשוואה):

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0) \quad \hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

ניתן לכתוב זאת במפורשת למשל עבור β_1 :

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$



רגרסיה לינארית פשוטה

<u>הערכת הדיוק של מקדמי המשוואה</u>

אם נחזור לדוגמא שלנו נקבל רווחי סמך עבור המקדמים β_0 ו - β_1: $[6.130, 7.935]$, $[0.042, 0.053]$

<u>המשמעות:</u> בהינתן שלא נפרסם בכלל המכירות ינועו בין 6130-7940, כאשר על כל פרסום נוסף ב1000$ נמכור בין 42-53 יותר יחידות

<u>בדיקת השערות</u>

זהו תחום בסטטיסטיקה בו אנו בוחנים השערה מסויימת H_1 על התפלגות הנתונים, אל מול ההנחה הבסיסית H_0.

<u>לדוגמא:</u> H_0- אין קשר בין המשתנה X ל-Y

H_1- ישנו קשר בין המשתנה X ל-Y

<u>או בצורה מתמטית:</u>

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

רגרסיה לינארית פשוטה

<u>הערכת הדיוק של מקדמי המשוואה</u>

אנו מעוניינים להחליט האם המשתנה שקיבלנו β_1 "מספיק רחוק" מ-0 על מנת שנוכל לקבוע כי יש קשר בין X ל- Y (β_1 הוא המקדם של "עוצמת" הקשר)

<u>כמה רחוק זה "מספיק רחוק"?</u>

זה כמובן תלוי בסטיית התקן. ככל שסטיית התקן קטנה (אנו יותר בטוחים בערכים שקיבלנו) אפילו אם נראה ערכים שרחוקים קצת מ-0 נקבע שיש קשר. לעומת זאת, אם סטיית התקן גדולה נצטרך ערכים רחוקים יותר על מנת לספק קביעה דומה.

בפועל אנו מחשבים סטטיסטי (מדד) שמחשב את מס' סטיות התקן בהן רחוק β_1 מ-0:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

רגרסיה לינארית פשוטה

<u>הערכת הדיוק של מקדמי המשוואה</u>

אם אין קשר בין המשתנים אנו נקבל התפלגות t (מעל 30 דגימות בערך דומה לנורמלית), בהתאם נוכל לחשב את ההסתברות לקבל את הערך המחושב של הסטטיסטי t (תחת ההנחה שאכן <u>אין קשר</u> בין X ל-Y)

<u>במילים אחרות:</u>

1. מצאנו אמד למשתנה מסוים
2. קיבלנו עבורו סטיית תקן
3. ניתן לייצג אותו (לפי השערת ה-0) כמשתנה t
4. ניתן לראות **כמה חריג** הערך שהתקבל. אם למשל הערך שקיבלנו חריג יותר מ95% מהערכים האפשריים (כלל ההתפלגות) נאמר שכנראה המשתנה אכן שונה מ-0 ו-**יש קשר** (או השערת ה-0 נדחית)

ההסתברות לקבל ערך כפי שקיבלנו (או קיצוני ממנו) נקרא : $\underline{\boldsymbol{P\_value}}$.

ככל שה- P_value קטן יותר משמע שהסיכוי לקבל כזה ערך (או קיצוני ממנו) בהנחה שבאמת אין קשר הוא מאוד נמוך, וכתוצאה מכך נחליט שישנו קשר

רגרסיה לינארית פשוטה

הערכת הדיוק של מקדמי המשוואה

בחזרה אלינו :

אם נבצע את הניתוח הנ"ל על הנתונים שלנו נקבל:

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

מסקנות:

1. שני המשתנים β_0 ו - β_1 (החותך ופרסום בטלוויזיה) קשורים למשתנה Y (מכירות)
2. למרות שהערכים של שני המשתנים אינם קרובים ניתן לראות שערכי סטיית התקן שלהם מותאמים וכתוצאה מכך הסטטיסטי דומה.

<u>הערכת הדיוק של המודל</u>

לאחר שדחינו את השערת ה-0 וראינו שישנו קשר בין המשתנים שלנו למשתנה המבוקש (Y) הצעד הבא יהיה להעריך את טיב המודל שלנו.

<u>לצורך הערכת טיב המודל ישנם שני מדדים מרכזיים:</u> RSE ,R^2

<u>RSE</u>

ניתן להתייחס למדד זה כאל הערכה לסטיית התקן של משתנה השגיאה ϵ , ובניסוח אחר ממוצע הסטייה מקו הרגרסיה:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

כמובן, ככל שהמדד יותר גדול כך הסטייה גדולה והמודל <u>פחות</u> מתאים לנתונים שבידינו

רגרסיה לינארית פשוטה

<u>הערכת הדיוק של המודל</u>

<u>$\boldsymbol{R^2}$</u>

RSE אומנם מודד את טיב המודל מבחינת ההתאמה לY (כפי שנראה בהמשך) אבל אינו מספק מדד יחסי אלא <u>תלוי בערכי Y</u>.

לצורך כך ישנו מדד R^2 שתפקידו לבדוק את : <u>כמות השונות המוסברת</u> (נע בין 0 ל-1) והוא אינו תלוי בסקאלה של Y.

<u>ניתן לפרמל זאת כך</u>

$$\mathrm{TSS} = \sum(y_i - \bar{y})^2,\ R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

נשים לב ש TSS מגדיר כמה שונות מובנה בתוצאות עצמן (כמה הן מגוונות) עוד לפני שהתאמנו את קו הרגרסיה, ואילו RSS מגדיר את השונות שנותרה לאחר מכן. כתוצאה מכך אנו מקבלים את השונות המוסברת.

<u>ככל שהמדד יותר גבוה כך יותר שונות מוסברת והמודל מתאים יותר טוב ולהפך</u>. אם נקבל מדד מאוד נמוך (קרוב ל-0) יהיה ניתן להסביר זאת בשונות מובנה גבוהה (TSS) או בחוסר התאמה של מודל רגרסיה לינארי (RSS) לייצוג המודל.

נזכור שR^2 הוא מדד לקשר הלינארי בין X לבין Y . ישנו מדד דומה שנקרא קורלציה :

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

נגדיר את הקורלציה כ-r , ניתן להראות כי $R^2 = r^2$ אך זאת במקרה של רגרסיה לינארית פשוטה בלבד.

רגרסיה לינארית פשוטה

הערכת הדיוק של המודל

המחשה:

Quantity	Value
Residual standard error	3.26
R^2	0.612

נחזור לדוגמא שלנו, נבצע הערכה של שני המדדים שהצענו ונקבל:

אז מה זה אומר לנו?

1. RSE– הטעות הממוצעת בהערכה של השגיאה היא 3.26 ולכן, על כל תחזית שנבצע (עם משתנה הטלוויזיה) נטעה בממוצע ב3260 יח' (האם זה מספיק טוב?) אם נסתכל על הנתונים נגלה שממוצע המכירות הוא: 14,000 כך שאחוז השגיאה הוא: $3{,}260/14{,}000 = 23\,\%$
2. R^2- השונות המוסברת היא 0.612 כלומר מתחת לשני שליש מהשונות הכללית של המודל.

האם המדדים הללו מספקים אותנו?

רגרסיה לינארית מרובה

<u>הקדמה</u>

כעת, נניח כי אנו רוצים לחזות את ערך המכירות על בסיס מספר פרמטרים שונים.

<u>נסיון ראשון:</u> יצירת קו רגרסיה עבור כל פרמטר בנפרד

האם זהו רעיון טוב? למה?

<u>נסיון שני:</u> יצירת קו רגרסיה אחד במספר מימדים:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

כאשר לכל פרמטר ישנו מקדם β, אותו אנו צריכים ללמוד (וכמובן חותך ומשתנה המייצג את השגיאה הבלתי ידועה) ובמקרה של הדוגמא שלנו זה ייראה כך:

$$\mathtt{sales} = \beta_0 + \beta_1 \times \mathtt{TV} + \beta_2 \times \mathtt{radio} + \beta_3 \times \mathtt{newspaper} + \epsilon$$

רגרסיה לינארית מרובה

<u>הערכת הפרמטרים</u>

גם כאן, כמו ברגרסיה פשוטה ישנה נוסחה סגורה לחישוב הפרמטרים (גזירת ה-Least squares)

לצורך כך נראה מעט אלגברה.(ללא התעמקות אלא בקווים כלליים)

ניתן לייצג את המודל שראינו מקודם בצורה מטריציונית כך: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

<u>כאשר</u> : Y- וקטור של התיוג/ערך כמספר התצפיות (N- מספר תצפיות)

X- מטריצת התצפיות [$N \times P$, N- מס' תצפיות , P – מספר משתנים]

β- מטריצת המקדמים –וקטור בגודל מספר המקדמים (P – פרמטרים)

$\mathcal{E}$- וקטור של השגיאות (בגודל התצפיות)

רגרסיה לינארית מרובה

הערכת הפרמטרים

לאחר קצת התעסקות מתמטית נוכל לקבל את הביטוי הבא: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

זוהי בעצם מטריצת המקדמים המשוערכים על פי הנתונים שבידינו.

נציין בהמשך לרגרסיה הפשוטה את השונות המשותפת של השערוך למקדמים: $Cov(\hat{\beta}) = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}$

מכאן ניתן להסיק את רווחי הסמך עבור הפרמטרים:

$$\widehat{Cov}(\hat{\beta}) = \hat{\sigma}^2 (\mathbf{X}^T\mathbf{X})^{-1} \longrightarrow \left(\hat{\beta}_j - 2\left\{\widehat{Cov}(\hat{\beta})\right\}_{jj}^{1/2}, \hat{\beta}_j + 2\left\{\widehat{Cov}(\hat{\beta})\right\}_{jj}^{1/2}\right)$$

גם כאן, בדומה לגרסיה פשוטה, ניתן לבצע מבחן t לבדיקת עוצמת הראיה לקשר בין הפרמטר לתוצאה כאשר נקבל את ה- ***P_value*** – ההסתברות לקבל ערך כזה של מקדם (כפי שקיבלנו) בהנחה שאין באמת קשר בין השניים (ככל שיותר קטן יותר טוב – בד"כ מתחת ל0.05 = מובהק)

רגרסיה לינארית מרובה

הערכת הפרמטרים

בדוגמא שלנו:

רגרסיה פשוטה

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

רגרסיה מרובה

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

ניתן לראות שישנו הבדל בין מובהקות הקשר של פרסום בעיתון למכירות בין רגרסיה פשוטה לבין רגרסיה מרובה.

<u>ניתן להסביר את הקשר הזה באמצעות מטריצת הקשרים בין המשתנים:</u>

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

אפשר לראות שהקשר בין מכירות לפרסום בעיתון עובר דרך <u>פרסום ברדיו</u>

רגרסיה לינארית מרובה

מבחן לבדיקת רגרסיה מרובה

בנוסף למבחן t עבור כל פרמטר נוכל לבצע מבחן נוסף:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

זהו מבחן לבדיקת ההשערה שישנו קשר (כלשהו) בין המשתנים לבין הערך החזוי.

זהו מבחן הנקרא **מבחן F:**

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$\mathrm{TSS} = \sum(y_i - \bar{y})^2$$

ככל שהסטטיסטי גדול יותר כך ההסתברות יותר גבוהה שישנו קשר בין (לפחות) אחד המשתנים לבין הערך החזוי. **מצד שני** ככל שהערך קרוב יותר ל-1 (אינו יכול להיות נמוך יותר) כך קטנה ההסתברות שישנו קשר.

שימוש נוסף למבחן זה הוא בהשוואה בין מודלים שונים (מודל עם q פרמטרים מול מודל עם $n-q$ פרמטרים)

אם יש לנו את מבחן t עבור כל פרמטר מדוע אנו זקוקים למבחן F ?

רגרסיה לינארית מרובה

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$$

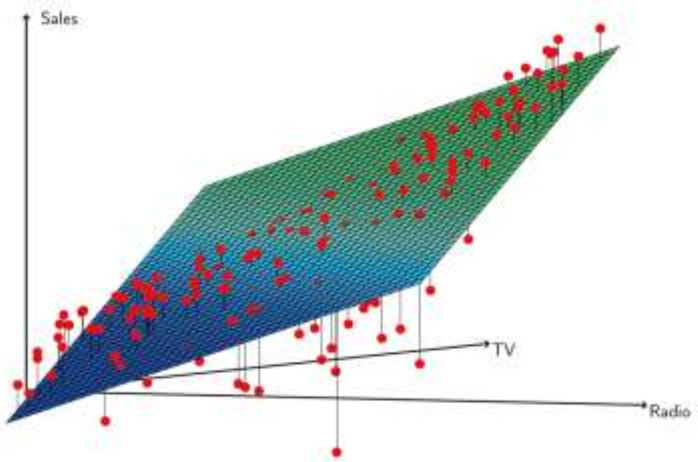
$$R^2 = \frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

<ins>הערכת טיב המודל</ins>

בדומה לרגרסיה פשוטה, יש לנו שני מדדים מרכזיים R^2, RSE .

<ins>בהקשר למדדים אלו חשוב לציין:</ins>

1. R^2 יעלה ככל שנוסיף פרמטרים (ללא תלות האם אם באמת קשורים למשתנה החזוי),לכן אנו נבחן באמצעות ההפרש בין הסטטיסטי המתקבל ללא הפרמטר ואיתו את השיפור ונוכל להחליט בהתאם.
2. RSE יכול לעלות (עליה -> יותר טעויות) על אף שנוסיף משתנים זאת כיוון שאומנם הטעות (במדגם האימון כמובן) תרד אבל מצד שני מספר הפרמטרים יורד גם הוא (ולכן זה תלוי ביחס)
3. ניתן לצייר את המישור הנוצר באמצעות הרגרסיה המרובה וללמוד ממנו על הנתונים



<ins>מה תוכלו להסיק מהפלט מצד ימין?</ins>

הרחבות למודל הלינארי

בעיות המודל הלינארי

ישנן שתי בעיות מרכזיות במודלים שהצגנו:

- בעיית האי-תלות- הנחנו כי המשתנים אינם תלויים בינם לבין עצמם, כלומר שינוי במשתנה מסוים משפיע על הערך החזוי ללא קשר למשתנים אחרים
- בעיית הלינאריות- הנחנו כי המודל מתנהג בצורה לינארית, כלומר שינוי הערך החזוי בתגובה לשינוי יחידה אחת במשתנה מסוים היא קבועה

אלו הנחות מפליגות ובפועל אינן מתקיימות. כיצד נוכל להתמודד איתן?

הרחבות למודל הלינארי

בעיות האי-תלות

נפתור את הבעיה באמצעות יצירת משתנה חדש הקושר בין המשתנים ומייצג את התלות ביניהם:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

כאשר ניתן לייצג זאת כך:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

כיצד המשתנה החדש פותר את הבעיה:

מכיוון שכעת $\tilde{\beta}_1$ מורכב גם מ-X_2 אז רמת ההשפעה של X_1 על Y תלויה גם בו $\Leftarrow$ פתרנו את הבעיה

הרחבות למודל הלינארי

<u>בעיית האי-תלות</u>

<u>אם נסתכל על הדוגמא שלנו:</u>

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

ניתן לראות שלמשתנה החדש שלנו בהחלט ישנו קשר מובהק לערך החזוי, נוכל להסביר זאת שעל כל עלייה בתקציב הפרסום של טלוויזיה המכירות יעלו ב- $19 + 1.1 \times radio$

מעבר לכך, R^2 עבור המודל עם המשתנה המשותף -96.8%

R^2 עבור המודל ללא המשתנה המשותף -89.7%

המודל החדש מסביר 69% מהשונות שלא הוסברה ע"י המודל הבסיסי.

אלו בעיות יכולות להיות עם סוג פתרון כזה?

הרחבות למודל הלינארי

<u>בעיית הלינאריות</u>

נפתור את הבעיה באמצעות <u>רגרסיה פולינומיאלית</u> (זו שיטה אחת מיני רבות)

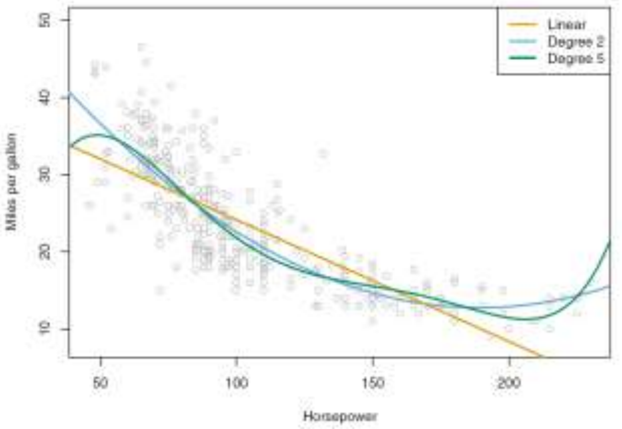
בתמונה מצד ימין מתואר מודל לינארי של צריכת דלק כפונקציה של כוח-סוס אם נסתכל על הנתונים נראה שהם מתנהגים בצורה לא לינארית, מה שאומר שרגרסיה פשוטה כנראה לא תתאים...

הפתרון שהמוצע הוא פשוט להכניס משתנים לא לינאריים לתוך המודל הלינארי:

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

זהו עדיין מודל לינארי! נוכל לראות שאכן כפי שצפינו הקשר למשתנה הריבועי הוא <u>מובהק</u>

<u>אולי כדאי להוסיף עוד משתנים?</u>



	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	−0.4662	0.0311	−15.0	< 0.0001
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

Gradient Descent

כאשר דיברנו על מודל הרגרסיה המרובה, הזכרנו את חישוב המקדמים: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

בפועל כאשר נרצה לחשב את וקטור המקדמים נבחין כי זוהי פעולה מאוד כבדה מבחינה חישובית.

לצורך ההמחשה נניח ובידינו מטריצת X בעלת K פרמטרים וN דגימות . אם K =1000 ו-N =1000000 אזי חישוב המטריצה ההופכית הופך להיות משימה חישובית כבדה ביותר.

לכן, אנו נוהגים להשתמש באלגוריתם <u>Gradient Descent :</u>
זהו אלגוריתם איטרטיבי שמטרתו למצוא את מקדמי המשוואה כך שיורידו למינימום את שגיאת האימון.

1. האלגוריתם מקבל את שגיאת האימון בסבב הנוכחי, המקדמים הנוכחיים וצעד בגודל α
2. גוזר אותה לפי כל אחד מהפרמטרים (נגזרת חלקית) ומוצא את הגרדיאנט (כיוון הירידה המקסימלי)
3. מעדכן בהתאם לגרדיאנט ולצעד את המקדמים
4. מבצע סבב נוסף וחוזר חלילה.

Gradient Descent

Objective:

$$\min_{\theta_0,\,\theta_1} J(\theta_0, \theta_1)$$

(1)

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Derivatives:

(2)

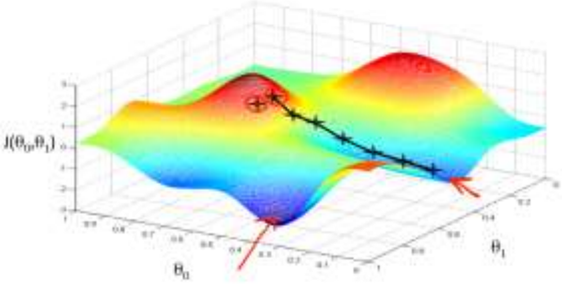
$$\frac{d}{d\theta_0}J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$\frac{d}{d\theta_1}J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)\cdot x^{(i)}$$

(3) Update rules:

$$\theta_0 := \theta_0 - \alpha\frac{d}{d\theta_0}J(\theta_0, \theta_1)$$

$$\theta_1 := \theta_1 - \alpha\frac{d}{d\theta_1}J(\theta_0, \theta_1)$$



שאלות להרחבה על המודל הלינארי

כפי שהדגשנו לאורך כל הדרך, המודל הלינארי הוא פשוט אך מוגבל במידת מה ולכן עלולות להתעורר בו בעיות שונות שעל חלקן ענינו ואת חלקן נשאיר פתוחות:

1. התמודדות עם נתונים שאינם לינאריים- ראינו שבד"כ הנתונים אינם מתנהגים בצורה לינארית ולכן עלינו לפתח שיטות לבדוק זאת ולהתמודד בהתאם
2. קורלציה בין משתני השגיאה (ϵ) -אנחנו הנחנו שהשגיאות אינן תלויות, לנתון זה ישנן השלכות משמעותיות במקרה ואיננו מדויק
3. שונות שאינה קבועה למשתני השגיאה- אנו הנחנו $\mathrm{Var}(\epsilon_i) = \sigma^2$, בפועל הנחה זו אינה תמיד מתקיימת ויוצרת בעיות במבחני ההשוואות
4. התמודדות עם outliers- יכולים להשפיע דרמטית על הערכת השגיאה והסברת השונות, כיצד אנו צריכים להתמודד איתם?
5. קו-לינאריות – לעיתים ישנם מספר משתנים שקשורים אחד לשני ומתנהגים בהתאם. יהיה קשה לנו לבודד משתנה אחד ולהעריך את התרומה שלו למודל.

סיכום

כעת, לאחר שיש לנו קצת יותר ידע בתור סטטיסטיקאים נחזור לשאלות שפתחנו איתן וננסה לתת תשובות בהתאם לחומר שלמדנו:

- <u>האם יש כלל קשר בין פרסום (כלשהו) לבין המכירות ? (אולי לא כדאי בכלל לפרסם?)</u>

נבנה מודל רגרסיה מרובה ונבצע מבחן F לבדיקת ההשערה האם לאחד (לפחות) מהמשתנים השונים (פרסום ברדיו, עיתון וטלוויזיה) ישנה השפעה.

נוכל לראות שה P_value שיתקבל ממבחן זה הוא מאוד נמוך מה שמצביע על <u>קשר כמעט וודאי</u>

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

- <u>בהנחה שישנו קשר בין תקציב הפרסום לבין המכירות (והתקציב ידוע), האם אנו יכולים לחזות (בדיוק גבוה) את היקף המכירות?</u>

נבחן את מודל הרגרסיה שהתקבל ונבדוק את הסטטסטיים שהצענו $(RSE\ ,R^2)$, לבחינת השונות המוסברת, במידה ונקבל שהמודל אכן מסביר בצורה טובה הנתונים נסיק שנוכל לחזות בדיוק גבוה את התוצאות. אם אכן נעשה זאת נגלה שקיבלנו משתנים המצביעים על התאמה טובה של המודל : $RSE = 1681\ ,R^2 \approx 90\%$ ⟸ $precentage\ error = 12\%$

סיכום

- איזו מדיה תורמת להגדלת היקף המכירות? האם כולן? האם אנו יודעים להעריך זאת ברמת דיוק גבוהה?

לאחר שנתאים מודל רגרסיה נוכל לבצע מבחן t פשוט עבור כל המשתנים ולראות מי תורם באופן מובהק. נוכל גם בנוסף להעריך כמה משפיע שינוי של יחידה אחת בפרמטרים אלו על ערך המכירות. במקרה שלנו קיבלנו כי המשתנים טלוויזיה ורדיו הם בעלי השפעה

- האם הקשר בין הפרסום לבין המכירות הוא לינארי?

באמצעות פלט של הנתונים על משטח הבחנו שהקשר איננו לינארי (ולטפל בנקודה זו בהתאם להסרת ההנחה הלינארית)

- האם ישנו קשר בין אמצעי הפרסום השונים? (למשל: יש לי תקציב של 100K$ כיצד לחלק אותו נכון ?)

ניתן לבדוק זאת באמצעות המשתנה המקשר (X_1X_2) ובחינת המובהקות הסטטיסטית שלו. אנו ראינו כי המשתנה הוא בעל מובהקות סטטיסטית ולכן ישנו קשר בין המשתנים טלוויזיה ורדיו

סיכום

<u>מה ראינו היום?</u>

- שיטות שונות להערכת השגיאה
- Bias vs Variance
- רגרסיה לינארית פשוטה
  - סטטיסטיקה בסיסית
  - בחירת מקדמים והערכת הדיוק שלהם
  - הערכת דיוק המודל
- רגרסיה מרובה
- הצגת אלגוריתם Gradient Descent

משימות לפגישה הבאה

- היכרות עם פלטפורמת DataCamp
  - ביצוע קורס "Intro to python for Data-Science" (עדיף גם את קורס 'Intermediate python for Data-Science")
  - ביצוע קורס "pandas foundation"
  - ישנם קורסים רבים תחת שרשרת קורסים: Data Scientist with Python רצוי לעשות קורסים ככל שניתן מהקורסים הבסיסיים (toolbox , Cleaning , Importing ועוד)
- צפייה בהדגמה של Gradient Descent
  - https://www.youtube.com/watch?v=yFPLyDwVifc- קורס Machine Learning מועבר ע"י Andrew Ng

תודה על ההקשבה