רגרסיה ליניארית מרובה

רגרסיה ליניארית מרובה הינה הרחבה של מודל רגרסיה ליניארית פשוטה (רל״פ).

במקום משתנה מסביר אחד בלבד, ברגרסיה מרובה יש מספר רב של משתנים מסבירים: $x_1,.x_2,\ldots,x_k$

. למודל רגרסיה מרובה יישומים רבים כאשר מעוניינים להעריך את ההשפעה הסימולטנית של מספר גורמים על משתנה תגובה כלשהו (המשתנה המוסבר).

נקבל אוסף של n תצפיות: $(y_1,x_{11},\ldots,x_{1k}),(y_2,x_{21},\ldots,x_{2k}),\ldots,(y_n,x_{n1},\ldots,x_{nk})$,

כאשר y הוא המשתנה התלוי.

לדוגמה: רוצים לבחון את הפרמטרים המשפיעים על מחיר דירה.

המשתנה התלוי הוא מחיר הדירה.

המשתנים המסבירים הם מספר החדרים בדירה, שטחה, והקומה שהיא נמצאת בה.

מחיר הדירה מושפע מהקומבינציה של המשתנים המסבירים (שילוב של מספר חדרים, שטח וקומה).

Yi	x3	x2	x1	i
מחיר ב-K$	קומה	שטח	חדרים	דירה
191	3	80	2	1
391	10	117	4	2
139	2	89	2	3
275	3	100	4	4
235	4	89	3.5	5
363	8	155	6	6
327	5	123	4.5	7
408	5	111	4.5	8
395	9	122	5	9
474	8	166	7.5	10
…	…	…	…	…

<u>מודל הרגרסיה הליניארית המרובה</u>

הנחות מודל הרגרסיה הליניארית המרובה מהוות הכללה של ההנחות שראינו עבור רל״פ. בפרט:

y_i - התצפית ה-i של המשתנה המוסבר

$-\varepsilon_i \sim N\left(0,\sigma^2\right)$ רעש אקראי. שימו לב שהשונויות שוות לכל i.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i$$

$$E(Y|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$\beta_0,\beta_1,\beta_2,\ldots,\beta_k$ הם מקדמי הרגרסיה, כאשר:

- β_0 הוא החותך - התוחלת של Y כאשר כל המשתנים המסבירים שווים ל-0.
- β_i הוא השיפוע של מישור הרגרסיה בכיוון של x_i, כלומר התוספת לתוחלת של Y כתוצאה מהגדלת X_i ביחידה אחת, כאשר שאר המשתנים מוחזקים קבועים

שאלה:

נתונה טבלת מספרים עם משתנים מנבאים $X_1, X_2 \ldots X_n, Y, n = 20$

בהינתן שני מודלי רגרסיה שונים:

1. מודל רגרסיה מלא המכיל את כל המשתנים המסבירים.
2. מודל רגרסיה חלקי $X_1, X_2 \ldots X_{10}, Y$

עבור המשתנים המסבירים הנמצאים בשני המודלים, האם האמדים (השיפוע) שלהם זהה?

<u>בדיקת מובהקות מודל הרגרסיה השלם – מבחן F</u>

היינו רוצים לבדוק אם המשתנים המסבירים $\left(x_1,\ldots,x_k\right)$ משפיעים על המשתנה התלוי (y).

מספיק קשר עם משתנה מסביר אחד (עם אחד מה – x – ים) כדי להגיד שקיימת השפעה כזו!

פורמלית, מערכת ההשערות שלנו היא:

$$\begin{cases} H_o: \beta_i = 0 \ \forall i \\ H_1: else \end{cases} \Rightarrow \begin{cases} H_o: \dfrac{E(MS_R)}{E(MS_e)} = 1 \\ H_1: \dfrac{E(MS_R)}{E(MS_e)} > 1 \end{cases}$$

סטטיסטי המבחן, כמו ברל"פ, הוא $F_0 = \frac{MSR}{MSE}$.

SST, SSE, SSR מוגדרים באופן דומה לרל"פ, אבל מספר דרגות החופש שלהם שונה:

Source of Variation	**Sum of Squares**	$\boldsymbol{d.f.}$	**Mean Square**	$\boldsymbol{F_0}$
Regression	$\boldsymbol{SSR} \equiv \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	k	$\boldsymbol{MSR} = \frac{SSR}{k}$	$F_0 = \frac{MSR}{MSE}$
Error	$\boldsymbol{SSE} \equiv \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$	$n-k-1$	$\boldsymbol{MSE} = \frac{SSE}{n-k-1}$	
Total	$\boldsymbol{SST} \equiv SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

לכן **כלל ההכרעה בר"מ** $\boldsymbol{\alpha}$: דחה אם $F_0 > F_{1-\alpha}^{(k,n-k-1)}$

כמו ברל"פ, MSE מהווה אומד חסר-הטייה ל-σ^2 (שונות הרעש).

<u>בעיה לדוגמא – איי גלפגוס</u>

מעוניינים למצוא את הגורמים המשפיעים על מספר זני בעלי החיים בכל אחד מהאיים בקבוצת איי גלפגוס. הגורמים שהוצעו הינם:

Area – שטח האי (קמ"ר)

Elevation – גובה הנקודה הגבוהה ביותר באי (מטרים)

Nearest – מרחק האי הקרוב ביותר (ק"מ)

Scruz – המרחק מהאי Santa Cruz (ק"מ)

Adjacent – שטח האי הקרוב ביותר (קמ"ר)

Species – מספר הזנים שניתן למצוא על האי.

נאספו נתונים על 30 איים, מרוכזים בטבלה בעמוד הבא.

island	Species	Area	Elevation	Nearest	Scruz	Adjacent
i	Yi	X1	X2	X3	X4	X5
Baltra	58	25.09	346	0.6	0.6	1.84
Bartolome	31	1.24	109	0.6	26.3	572.33
Caldwell	3	0.21	114	2.8	58.7	0.78
Champion	25	0.1	46	1.9	47.4	0.18
Coamano	2	0.05	77	1.9	1.9	903.82
Daphne.Major	18	0.34	119	8	8	1.84
Daphne.Minor	24	0.08	93	6	12	0.34
Darwin	10	2.33	168	34.1	290.2	2.85
Eden	8	0.03	71	0.4	0.4	17.95
Enderby	2	0.18	112	2.6	50.2	0.1
Espanola	97	58.27	198	1.1	88.3	0.57
Fernandina	93	634.49	1494	4.3	95.3	4669.32
Gardner1	58	0.57	49	1.1	93.1	58.27
Gardner2	5	0.78	227	4.6	62.2	0.21
Genovesa	40	17.35	76	47.4	92.2	129.49
Isabela	347	4669.32	1707	0.7	28.1	634.49
Marchena	51	129.49	343	29.1	85.9	59.56
Onslow	2	0.01	25	3.3	45.9	0.1
Pinta	104	59.56	777	29.1	119.6	129.49
Pinzon	108	17.95	458	10.7	10.7	0.03
Las.Plazas	12	0.23	94	0.5	0.6	25.09
Rabida	70	4.89	367	4.4	24.4	572.33
SanCristobal	280	551.62	716	45.2	66.6	0.57
SanSalvador	237	572.33	906	0.2	19.8	4.89
SantaCruz	444	903.82	864	0.6	0	0.52
SantaFe	62	24.08	259	16.5	16.5	0.52
SantaMaria	285	170.92	640	2.6	49.2	0.1
Seymour	44	1.84	147	0.6	9.6	25.09
Tortuga	16	1.24	186	6.8	50.9	17.95
Wolf	21	2.85	253	34.1	254.7	2.33

פלט הרגרסיה:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8751
R Square	0.7658
Adjusted R Square	0.7171
Standard Error	60.9752
Observations	30

מבחן F לכל המודל

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	291850.0003	58370.0001	15.6994	6.83789E-07
Residual	24	89231.3663	3717.9736		
Total	29	381081.3667			

מבחן t לכל מקדם בנפרד

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	7.0682	19.1542	0.3690	0.7154	-32.4641	46.6005
Area	-0.0239	0.0224	-1.0676	0.2963	-0.0702	0.0223
Elevation	0.3195	0.0537	5.9532	0.0000	0.2087	0.4302
Nearest	0.0091	1.0541	0.0087	0.9932	-2.1665	2.1848
Scruz	-0.2405	0.2154	-1.1166	0.2752	-0.6851	0.2040
Adjacent	-0.0748	0.0177	-4.2262	0.0003	-0.1113	-0.0383

$$F_{Cr}=F_{0.05}(5,24)=2.62 \quad < \quad F_0=15.6994$$

נבדוק את כלל הדחייה: $F_0=15.6994>F_{cr}=F_{0.95}^{(5,24)}=2.62$ ולכן נדחה את H_0.
כלומר: מספר הזנים השונים על איי הגלפגוס **<u>כן</u>** מושפע מלפחות אחד מהמשתנים שבדקנו.
שיקול נוסף שאפשר להפעיל: PV של סטטיסטי המבחן נמוך מאוד (סדר גודל של 10^{-7}) ולכן לכל רמת מובהקות סבירה (למשל, 5%) נדחה את השערת האפס.

<u>הסקה על מקדמי הרגרסיה</u>

לאחר מבחן כולל למובהקות המודל הרב משתני, נרצה לבדוק השערות לגבי פרמטרים בודדים או קבוצות של פרמטרים. **קיימות שתי גישות לבדיקת השערות של המקדמים: מבחן t, ומבחן F חלקי.**

(1) **<u>מבחן t:</u>** זהה למבחן המבוצע ברגרסיה פשוטה. מתבסס על ההתפלגות הנורמלית של אמדי הריבועים הפחותים ועל אמדי סטיות התקן המתקבלים בתהליך האמידה. מאפשר בחינה של כל פרמטר בנפרד.

אם הפרמטר ה-j שווה ל-0, אין למשתנה ה-j השפעה ליניארית על המשתנה המוסבר.

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

סטטיסטי המבחן הוא $t_0=\frac{b_j}{s(b_j)}$, כאשר $s(b_j)$ היא סטיית התקן של האומד למקדם של המשתנה ה-j (ערך זה מופיע בפלט של הרגרסיה).

כלל ההכרעה: דחה אם $|t_0|>t_{1-\frac{\alpha}{2}}^{(n-k-1)}$.

הערה (1): <u>רב"ס</u> דו"צ לפרמטר β_j ברמת סמך $1-\alpha$: $\beta_j\in[b_j\pm t_{1-\frac{\alpha}{2}}^{(n-k-1)}\cdot S(b_j)\ ]$

(2) **מבחן F חלקי** – בוחן את התרומה של קבוצת משתנים מסבירים להסבר שהרגרסיה מספקת (ניתן לבדוק פרמטר יחיד, או קבוצת פרמטרים בו"ז). משווים את SSR בלי קבוצת המשתנים, לעומת הערך של רגרסיה שכוללת אותם.

לצורך הדגמת הסימון שנשתמש בו, נניח שאנו בוחנים שלושה משתנים מסבירים - X_1,X_2,X_3.
$SSR(X_1,X_2,X_3)$ - השונות המוסברת ע"י מודל רגרסיה הכולל את שלושת המשתנים.
$SSR(X_2)$ – השונות המוסברת ע"י מודל רגרסיה הכולל רק את המשתנה X_2.
אז ניתן לומר ש"תוספת ההסבר" שהמשתנים X_1,X_3 מספקים היא:

$$SSR(X_1,X_2,X_3)-SSR(X_2)$$

נרצה לבדוק אם הערך הזה אכן <u>מובהק</u> בעזרת מבחן השערות.

באופן כללי, נגדיר:

המודל "המלא" – כולל קבוצת המשתנים שאנחנו בוחנים

המודל המצומצם – ללא קבוצת המשתנים שאנחנו בוחנים

בכל מודל, מחשבים בנפרד את SSR. מספר דרגות החופש $(d.f.)$ של SSR בכל מודל הוא מספר המשתנים המסבירים הנכללים בו.

סטטיסטי המבחן:
$$F_0 = \frac{(SSR_{full}-SSR_{partial})/(df_{full}-df_{partial})}{SSE_{full}/(n-df_{full}-1)}$$

שימו לב שניתן לחלק גם את המונה וגם את המכנה ב-SST, ולקבל את הביטוי החלופי הבא ל-F_0:
$$F_0 = \frac{({R^2}_{full} - {R^2}_{partial})/(df_{full} - df_{partial})}{(1 - {R^2}_{full})/(n - df_{full} - 1)}$$

<u>כלל ההכרעה ברמ"מ α:</u> דחה את השערת האפס אם $F_0 > F_{1-\alpha}^{(df_{full}-df_{partial},\ n-df_{full}-1)}$.

למשל, בדוגמת גלפאגוס:

נבדוק אם למשתנה Nearest יש תרומה מובהקת למודל.

פלט הרגרסיה ללא Nearest:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8751
R Square	0.7658
Adjusted R Square	0.7284
Standard Error	59.7433
Observations	30.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	291849.7206	72962.43014	20.44186	1.38984E-07
Residual	25	89231.64609	3569.265844		
Total	29	381081.3667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	7.075377	18.7498	0.3774	0.7091	-31.5406	45.6914
Area	-0.02398	0.0215	-1.1148	0.2756	-0.0683	0.0203
Elevation	0.319573	0.0511	6.2505	0.0000	0.2143	0.4249
Scruz	-0.23936	0.1646	-1.4538	0.1584	-0.5784	0.0997
Adjacent	-0.07485	0.0166	-4.5011	0.0001	-0.1091	-0.0406

במודל המלא (עמ' 3): $SSR_{full} = 291850$ $df_{full} = 5$

במודל החלקי (עמ' 5): $\mathrm{SSR}_{partial} = 291849$ $df_{partial} = 4$

$$F_0 = \frac{\frac{291850-291849}{5-4}}{\frac{89231.2663}{30-5-1}} = 0.00027 < 4.26 = F_{0.05}{}^{(1,24)} = F_{cr}$$

ולכן לא דוחים את H_0, כלומר **לפי מבחן F חלקי** למשתנה $Nearest$ אין השפעה ליניארית מובהקת (בר״מ 0.05) על מספר הזנים באי.

<u>נבדוק אם מגיעים לאותה מסקנה לפי מבחן t במודל הרגרסיה המלא (עמ׳ 3):</u>

כלל הדחייה של מבחן t: דחה אם $2.064 = t_{0.975}^{(24)} = t_{1-\frac{\alpha}{2}}^{(n-k-1)} \ngtr |t_0| = 0.0087$, ולכן לא נדחה את H_0- גם לפי מבחן t אין למשתנה המסביר $Nearest$ השפעה מובהקת על מספר הזנים באי.

ניתן להגיע למסקנה דומה גם בעזרת PV של המשתנה שהוא גבוה מאוד!

<u>מקדם ההסבר ברגרסיה ליניארית מרובה</u>

באופן כללי, ככל שמספר המשתנים המסבירים גדל, כך הרגרסיה יכולה להתאים משוואה טובה יותר לנתונים. לכן, מקדם ההסבר R^2 לעולם לא יירד כאשר נוסיף משתנים מסבירים למודל. לכאורה, סך השונות המוסברת גדל (וזה טוב), אולם הגידול הזה הינו טכני ומלאכותי, ולא נובע בהכרח מ״הסבר״ טוב יותר של המשתנים באופן מהותי.

לכן, שימוש ב-R^2 כמדד ליכולת ההסבר של המודל הינו בעייתי במודל רגרסיה ליניארית מרובה.

מדד חלופי הינו $R^2_{adj.}$, אשר כולל ״מעניש״ כנגד הכללת משתנים מסבירים רבים במודל:

n – מספר התצפיות
k – מספר המשתנים המסבירים במודל

$$R^2_{adj.} \equiv 1 - \frac{SSE/_{n-k-1}}{SST/_{n-1}}$$

ניתן להראות שבכל מודל רגרסיה מתקיים $R^2_{adj} \leq R^2$, כך שיש פחות תמריץ להוסיף הרבה משתנים למודל.

<u>בדוגמת גלפגוס:</u> ניתן לראות כי המודל ללא Nearest הינו בעל $R^2_{adj.}$ גבוה יותר מאשר המודל המלא, לכן נעדיף את המודל הזה. בשלב זה ניתן לנסות להוסיף/להוריד משתנים נוספים במטרה למצוא מודל טוב יותר.

דוגמה לחישוב $R^2_{adj.}$ במודל ללא Nearest (על סמך הפלט שמופיע בעמ׳ 5):

$$R^2_{adj.} = 1 - \frac{89231.646/_{25}}{381081.3667/_{29}} = 0.7284$$

<u>מולטיקולינאריות</u>

מולטיקולינאריות היא תופעה שבה קיים מתאם ליניארי חזק בין שני משתנים מסבירים (או יותר). במקרה כזה, המשתנים המסבירים מספקים מידע יתיר לגבי המשתנה המוסבר. מסתבר שלמולטיקולינאריות יש השלכות בעייתיות על האמינות והיציבות של האומדנים שהרגרסיה מספקת.

<u>**השלכות אפשריות של מולטיקולינאריות**</u> - אלו בד"כ סימנים מעידים לנוכחות של מולטיקולינאריות בנתונים:

(1) **האומדים לשונות של מקדמי הרגרסיה (s_{b_i}) "מתנפחים".** דבר זה גורם לכך שהרב"סים למקדמים האמיתיים (β_i) הרבה יותר רחבים מכפי שהם אמורים להיות, ואז הרגרסיה פחות אמינה.

(2) **תוצאות "מוזרות": סימני המקדמים הפוכים מהצפוי.**

(3) **תוצאות "מוזרות": מבחן F מעיד על מודל מובהק, בזמן שאף מבחן t אינו מובהק.**

<u>למה זה קורה?</u> מבחן F בודק אם המשתנה המוסבר מוסבר בצורה מספקת ע"י כלל המשתנים המסבירים. מבחן t לכל משתנה בנפרד בודק אם למשתנה זה יש תוספת הסבר מובהקת למשתנה המוסבר <u>כאשר כלל המשתנים האחרים כלולים כבר במודל.</u> כאשר המשתנים המסבירים מסבירים אחד את השני, ייתכן שהתרומה של כל אחד מהם בנפרד לא תהיה מובהקת, בעוד שבפועל הם אכן מסבירים בצורה טובה את המשתנה המוסבר.

<u>מבחנים למולטיקולינאריות</u>

<u>**דרך ראשונה - בדיקת המתאם בין כל זוג משתנים מסבירים בנתונים**</u>: האומד למתאם בין שני כל משתנים לפי מדגמים שלהם הוא $r_{xy} = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$.

באופן דומה, ניתן לחשב אומד למתאם בין כל שני משתנים מסבירים:

$$r_{x_j,x_k} = \frac{SS_{x_j,x_k}}{\sqrt{SS_{x_j} \cdot SS_{x_k}}}$$

אם המתאם גבוה (קרוב ל-1 בערך מוחלט), כדאי לוותר על אחד מהמשתנים הללו.

<u>**דרך שנייה – מדד VIF (Variance Inflation Factor):**</u> לכל משתנה מסביר x_j מחשבים את הערך $VIF_j = \frac{1}{1-R_j^2}$, כאשר R_j^2 הוא מקדם ההסבר עבור מודל רגרסיה שבו המשתנה המוסבר הוא x_j והמשתנים המסבירים הם שאר המשתנים המסבירים.

ערך מדד $VIF_j \geq 5$ מעיד על בעיה של מולטיקולינאריות (המשתנה <u>המסביר</u> x_j <u>מוסבר</u> ע"י שאר המשתנים מסבירים, כך שיש תיאום).

הערך של מדד VIF_j אומר <u>פי כמה</u> גדלה השונות של המקדם b_j בגלל התלות של המשתנה המסביר x_j במשתנים המסבירים האחרים.

<u>**דוגמה**</u>

מעוניינים לבנות מודל רגרסיה ליניארית שבו מסבירים את גובהו של אדם (y) באמצעות שני משתנים מסבירים – גודל כף הרגל הימנית (x_1), וגודל כף הרגל השמאלית (x_2). קובץ הנתונים שישרת אותנו בבעיה (מופיע באתר הקורס) הוא בעל המבנה הבא:

#	Right Foot (Inch)	Left Foot (Inch)	Height (Inch)
(1)	14.43	14.49	77.31
(2)	11.21	11.23	67.58
…			
(105)	12.02	12.10	69.57

נתחיל מניתוח תוצאות מודל רגרסיה מרובה שכולל את שני המשתנים:

Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0.9042	0.8176	0.8140	2.004141809

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	2	1836.384497	918.1922484	228.6003	< 0.0001
Unexplained	102	409.6916079	4.016584391		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	31.76029318	1.959464212	16.2087	< 0.0001	27.8737052	35.64688115
Right	6.822926303	3.428475129	1.9901	0.0493	0.02256214	13.62329047
Left	-3.644781741	3.441067666	-1.0592	0.2920	-10.47012314	3.180559658

הרגרסיה מסבירה היטב את הנתונים (לפי R^2 ו- R^2_{adj}.) והיא גם מובהקת לפי מבחן F. עם זאת, קיבלנו שתי תוצאות לא צפויות: המקדם של המשתנה x_2 שלילי (בניגוד לאינטואיציה שלנו), ומבחן t שלו אינו מובהק. זה צריך לעורר אצלנו נורה אדומה שישנה מולטיקוליניאריות בנתונים, לכן נבצע את הבדיקות המתאימות....

מבחן ראשון - מטריצת המתאם: בתא (i,j) מופיע r_{x_i,x_j}, האומדן למתאם בין x_i ו-x_j:

	גובה	גודל כף רגל ימין	גודל כף רגל שמאל
גובה	1		
גודל כף רגל ימין	0.903	1	
גודל כף רגל שמאל	0.900	0.999	1

שני המשתנים המסבירים מתואמים באופן (כמעט) מושלם.

מבחן שני – נחשב את ה-VIF של x_2: נבצע רגרסיה שבה המשתנה המוסבר הוא x_2 והמשתנה המסביר הוא x_1. ברגרסיה הזו $R^2 = 0.9981$ ואז $VIF_2 = \frac{1}{1-0.9981} = 525.19 \gg 5$. זו עדות נוספת לקיומה של מולטיקוליניאריות בנתונים.

אז מה עושים?

הפתרון הפשוט ביותר הוא להשמיט את המשתנה המסביר שתלוי במשתנים האחרים. במקרה הזה שני המשתנים תלויים אחד בשני, ואכן בכל פעם שמשמיטים אחד מהם מקבלים רגרסיה מובהקת.

הגובה כפונקציה של גודל כף רגל ימין :

Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0.9031	0.8156	0.8138	2.005327453

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	1	1831.878271	1831.878271	455.5395	< 0.0001
Unexplained	103	414.197834	4.021338194		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	31.5457001	1.950115191	16.1763	< 0.0001	27.67810656	35.41329363
Right	3.194941505	0.149692453	21.3434	< 0.0001	2.898061831	3.49182118

הגובה כפונקציה של גודל כף רגל שמאל :

Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0.9003	0.8105	0.8087	2.032739063

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	1	1820.477211	1820.477211	440.5772	< 0.0001
Unexplained	103	425.5988941	4.132028098		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	31.52588005	1.983829749	15.8914	< 0.0001	27.59142164	35.46033847
Left	3.19668202	0.152295983	20.9899	< 0.0001	2.894638858	3.498725182

בשתי הרגרסיות מקדם ההסבר גבוה, המודל כולו מובהק וגם מבחני t עבור המקדמים מובהקים.

ישנם גם פתרונות אפשריים נוספים למולטיקוליניאריות, אשר לא נכסה בקורס.

משתני דמי = משתנים קטגוריאליים

מודלי הרגרסיה שעסקנו בהם עד כה היו מבוססים על **משתנים כמותיים**, כלומר משתנים הנמדדים על ציר מספרי, לדוגמא: טמפרטורה, מרחק, גיל, עלות וכו'. לעיתים יש צורך בשילוב של משתנים איכותניים במודל הרגרסיה, לדוגמה שיוך לשכונה מסוימת.

כל ערך של משתנה כזה נקרא "רמה", והשיטה המקובלת להערכת ההשפעה של רמות שונות של משתנה איכותני על משתנה תלוי היא **שימוש באינדיקטורים**. באופן כללי, מייצגים משתנה איכותני בעל m רמות על ידי $m-1$ אינדיקטורים המקבלים את הערכים 0 או 1.

לדוגמה: מעוניינים לבדוק את רמת ההשכלה כמשתנה מסביר עבור השכר של בוגרי תואר בהנדסת תעשייה. זהו משתנה מסביר קטגוריאלי עם שלוש רמות: בוגר תואר ראשון, בוגר תואר שני, בוגר תואר שלישי.

על מנת למדל משתנה קטגוריאלי עם 3 רמות, נשתמש בשני משתני דמי x_1 ו-x_2, אותם נקודד באופן הבא:

בוגר תואר שלישי	בוגר תואר שני	בוגר תואר ראשון	
0	1	0	x_1
1	0	0	x_2

המודל לבחינת השפעת ההשכלה על השכר $y=\beta_0+\beta_1x_1+\beta_2x_2$:

תוחלת השכר של בוגרי תואר ראשון: $y_1=\beta_0$

תוחלת השכר של בוגרי תואר שני: $y_2=\beta_0+\beta_1$

תוחלת השכר של בוגרי תואר שלישי: $y_3=\beta_0+\beta_2$

המשתנה x_1 משמעו שצריך להוסיף "**קפיצה**" בגובה β_1 בערך של y כאשר מדובר על שכר של בוגרי תואר שני ביחס לבוגר תואר ראשון.
המשתנה x_2 משמעו שצריך להוסיף "**קפיצה**" בגובה β_2 בערך של y כאשר מדובר על שכר של בוגרי תואר שלישי ביחס לבוגר תואר ראשון.

אינטראקציות

אינטראקציה היא השפעה משולבת של מספר משתנים מסבירים. בכל המודלים שראינו עד עכשיו, הנחנו שגודל ההשפעה (=השיפוע) של כל משתנה מסביר אינו תלוי בערכים של המשתנים המסבירים האחרים. אבל לעתים מעניין לבדוק אם לשילוב של ערכים שונים של המשתנים המסבירים ישנה השפעה שונה, זוהי בדיקה של האינטראקציה בין המשתנים המסבירים.