פתרון תרגיל במתאם ורגרסיה

שאלה 1

בחינת ה-SAT היא המקבילה האמריקנית של הבחינה הפסיכומטרית בישראל. לבחינה יש שני חלקים, מילולי ומתמטי, ולכל חלק ניתן ציון נפרד.

בתיקיית הנתונים בדרייב של הקורס נמצא קובץ אקסל המכיל נתונים על 50 המדינות של ארצות הברית, וכן עבור מחוז קולומביה (Washington DC). כן תוכלו למצוא שם הסבר על מבנה הקובץ ומה מייצג כל משתנה.

בשאלה זו נבנה בעזרת אקסל מודל שיסביר את ציוני החלק המתמטי של המבחן (המשתנה satMath), וזאת על יד השכר הממוצע של המורים (המשתנה teacherPay) ואחוז הניגשים לבחינה (המשתנה percentTaking).

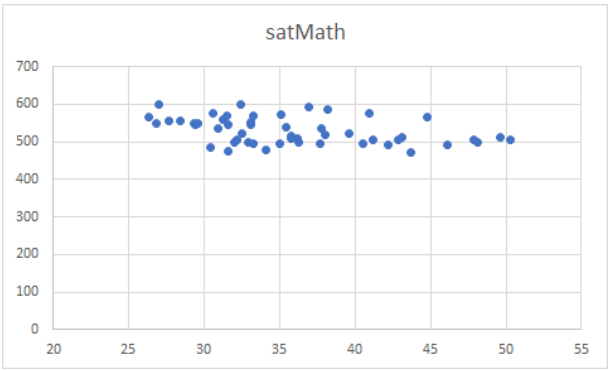
A	B	C	D	E	F	G	H
state	region	population	satVerbal	satMath	percentTaking	percentNoHS	teacherPay
AL	ESC	4273	565	558	8	33.1	31.3
AK	PAC	607	521	513	47	13.4	49.6
AZ	MTN	4428	525	521	28	21.3	32.5
AR	WSC	2510	566	550	6	33.7	29.3
CA	PAC	31878	495	511	45	23.8	43.1
CO	MTN	3823	536	538	30	15.6	35.4
CT	NE	3274	507	504	79	20.8	50.3
DE	SA	725	508	495	66	22.5	40.5

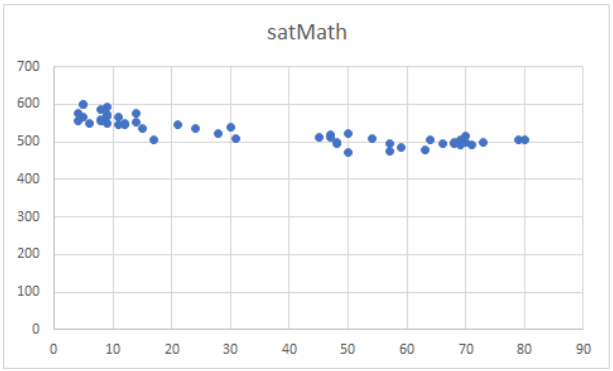
© Yossi Levy 2021

www.sci-princess.info

שאלה 1, סעיף א

א. צרו שתי דיאגרמות פיזור: אחת מהן תציג את הציון במתמטיקה מול שכר המורים, והשנייה תציג את הציון במתמטיקה מול אחוז הניגשים לבחינה. על סמך התבוננות בדיאגרמות, האם לדעתם ניתן להניח כי קיים קשר לינארי בין שכר המורים והציון במתמטיקה, וכן בין אחוז הניגשים לבחינה ובין הציון במתמטיקה?

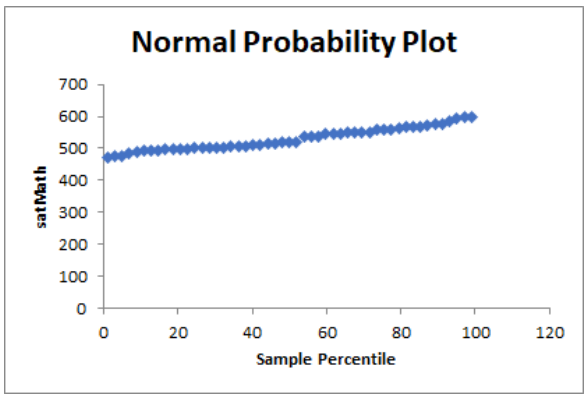




שאלה 1, סעיף ב

ב. צרו מודל רגרסיה בו המשתנה המוסבר הוא הציון במתמטיקה, והמשתנה המסביר הוא שכר המורים.

1) האם ההנחה כי התפלגות השאריות היא התפלגות נורמלית היא הנחה סבירה?



© Yossi Levy 2021

www.sci-princess.info

שאלה 1, סעיף ב

$satMath = 610.4 - 2.26\ teacherPay$

2) מהי משוואת הרגרסיה?
3) מהו אחוז השונות המוסברת?
4) הסבירו כיצד על פי המודל שכר המורים משפיע על הציון במתמטיקה.

ככל ששכר המורים עולה הציון במתמטיקה יורד!

Regression Statistics	
Multiple R	0.404
R Square	0.163
Adjusted R Square	0.146
Standard Error	32.189
Observations	51

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	610.40	26.63	22.92	0.0000	556.89	663.90
teacherPay	-2.26	0.73	-3.09	0.0033	-3.73	-0.79

שאלה 1, סעיף ב

5) אם במדינה מסויימת השכר הממוצע של המורים הוא 60 אלף דולר בשנה, מה צפוי להיות הציון במתמטיקה על פי המודל?
6) חוו את דעתכם על טיב המודל

הציון הצפוי על פי המודל הוא: $610.4 - 2.26 \cdot 60 = 474.8$

© Yossi Levy 2021
www.sci-princess.info

שאלה 1 סעיף ג

ג. צרו מודל רגרסיה בו המשתנה המוסבר הוא הציון במתמטיקה, והמשתנה המסביר הוא אחוז הניגשים לבחינה

$$satMath = 569.8 - 1.1 \cdot percentTaking$$

1) מהי משוואת הרגרסיה?
2) מהו אחוז השונות המוסברת?
4) האם דוחים את ההשערות כי מקדמי הרגרסיה (a ו-b) שונים מאפס?

Regression Statistics	
Multiple R	0.861
R Square	0.741
Adjusted R	0.736
Standard E	17.914
Observatic	51

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressio	1	44948.1	44948.1	140.1	0.0000
Residual	49	15724.0	320.9		
Total	50	60672.2			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	569.8	4.2	134.3	0.0000	561.2	578.3
percentTal	-1.1	0.1	-11.8	0.0000	-1.3	-0.9

שאלה 1 סעיף ג

3) מהו ערך השארית עבור מדינת אילינוי (קוד המדינה הוא IL)

הציון הממוצע במדינת אילינוי הוא 575, ואחוז הניגשים לבחינה הוא 14

הציון הצפוי על פי המודל הוא $satMath = 569.8 - 1.1 \cdot 14 = 554.4$

לכן שארית היא $575 - 554.4 = 20.6$

5) הסבירו כיצד על פי המודל אחוז הניגשים לבחינה משפיע על הציון במתמטיקה.
6) חוו את דעתכם על טיב המודל

$$satMath = 569.8 - 1.1 \cdot percentTaking$$

© Yossi Levy 2021

www.sci-princess.info

שאלה 1 סעיף ד

ד. צרו מודל רגרסיה בו המשתנה המוסבר הוא הציון במתמטיקה, הפעם עם שני משתנים מסבירים: שכר המורים ואחוז הניגשים לבחינה.

1) מהי משוואת הרגרסיה?
2) מהו אחוז השונות המוסברת?
3) מהו האמדן לסטיית התקן של השאריות?
4) האם דוחים את ההשערות כי מקדמי הרגרסיה שונים מאפס?

$$satMath = 537.9 - 1.3 \cdot \text{percentTaking} + 1.0 \cdot \text{teacherPay}$$

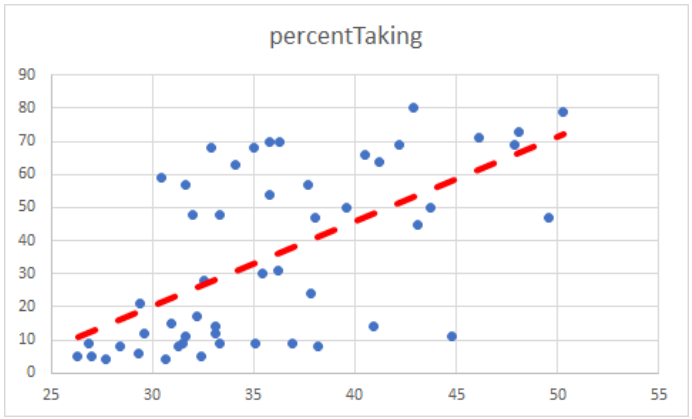
Regression Statistics	
Multiple R	0.873
R Square	0.762
Adjusted R Squar	0.753
Standard Error	17.3
Observations	51

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	537.9	15.8	34.1	0.0000	506.2	569.7
teacherPay	1.0	0.5	2.1	0.0421	0.0	2.0
percentTaking	-1.3	0.1	-11.0	0.0000	-1.5	-1.1

9

שאלה 1 סעיף ה

ה. מהו לדעתכם הסבר אפשרי לשוני בין מקדם הרגרסיה של שכר המורים במודל הראשון ובין המקדם המקביל במודל השלישי?



10

© Yossi Levy 2021

שאלה 2, סעיפים א-ב

שאלה 2

כדי לבדוק את הקשר בין המשתנה time: הזמן שסטודנט מתכונן למבחן (בשעות) ובין המשתנה score: הציון במבחן, חוקר אסף נתונים ובנה מודל רגרסיה. קו הרגרסיה הוא $score = 31.2 + 6.7 \cdot time$

א. מהו ניבוי קו הרגרסיה עבור סטודנט שהתכונן למבחן במשך 7 שעות?

$$score = 31.2 + 6.7 \cdot 7 = 78.1$$

ב. סטודנט התכונן למבחן במשך 5 שעות ולבסוף קיבל ציון 70. מהי השגיאה/שארית של תצפית זו?

הציון הצפוי על פי המודל הוא $score = 31.2 + 6.7 \cdot 5 = 64.7$

ולכן השגיאה היא $70 - 64.7 = 5.3$

שאלה 2, סעיף ג

ג. סטודנט תכנן ללמוד מספר מסויים של שעות, אז בסופו של דבר למד שעתיים פחות ממה שתכנן. מה יהיה השינוי בציון הצפוי לו על פי המודל?

נניח שתכנן ללמוד T שעות. הציון הצפוי הוא $score_T = 31.2 + 6.7 \cdot T$

לבסוף למד T-2 שעות ולכן הציון הצפוי הוא

$$score_{T-2} = 31.2 + 6.7 \cdot (T - 2) =$$

$$31.2 + 6.7 \cdot T - 6.7 \cdot 2 =$$

$$score_T - 6.7 \cdot 2 = score_T - 13.4$$

© Yossi Levy 2021

www.sci-princess.info

שאלה 2, סעיף ד

ד. לאחר סיום הניתוח הסטטיסטי, החוקר השיג נתון על סטודנט נוסף שהתכונן למבחן במשך 8 שעות. מהו הציון שקיבל סטודנט זה במבחן?

תשובה לא נכונה: $score = 31.2 + 6.7 \cdot 8 = 84.8 \approx 85$

מודל הרגרסיה נותן ניבוי/הערכה/ניחוש אינטליגנטי לציון הצפוי תחת הנחות מסויימות.

84.8 הוא חיזוי על פי המודל, לא הציון האמיתי.

התשובה הנכונה היא שאי אפשר לדעת את הציון על סמך הנתון של השאלה

שאלה 2 סעיף ה

ה. מה יהיה קו הרגרסיה אם נמדוד את הזמן בדקות במקום בשעות?

אם $time$ הוא הזמן בשעות אז הזמן בדקות הוא $minutes = 60 \cdot time$

$$6.7 \cdot time = \frac{1}{60} \cdot 60 \cdot 6.7 \cdot time = \frac{1}{60} \cdot 6.7 \cdot 60 \cdot time$$

$$= \frac{1}{60} \cdot 6.7 \cdot minutes = 0.112 \cdot minutes$$

ולכן משוואת הרגרסיה בדקות היא $score = 31.2 + 0.112 \cdot minutes$

© Yossi Levy 2021

www.sci-princess.info

שאלה 3

כדי לבדוק את הקשר בין איכות טעמו של יין ובין מרכיביו הכימיים נלקח מדגם מייצג של בקבוקי יין אדום. מכל בקבוק נלקחה דגימה לבדיקת מעבדה. כמו כן, צוות של עשרה מומחים טעם את כל היינות ודירג את איכותם. איכות היין היא הממוצע של הדירוג של עשרת המומחים.
לצורך תרגיל זה נתמקד בקשר בין איכות היין ובין רמת האלכוהול שבו, באחוזים. כמו כן, נתעלם מהשאלה האם ההנחות הדרושות מתקיימות.
כדי לבדוק את הקשר נבנה מודל רגרסיה. מצורף פלט תוצאת הרגרסיה מניתוח באקסל. למרבה הצער, המרצה של הקורס מחק ממנו כמה נתונים... עם זאת, נתון כי ממוצע דירוגי היינות הוא 5.61 וכי ממוצע רמת האלכוהול של היינות שבמדגם הוא 10.16.

שאלה 3, סעיפים א-ב

א. מהו המשתנה המסביר ומהו המשתנה המוסבר?

ב. מה היה גודל המדגם? $597 + 1 = 598$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R						
R Square						
Adjusted R Square	0.189					
Standard Error	0.764					
Observations						
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1		81.84			
Residual	596	347.79	0.58			
Total	597	429.63				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.916	0.31			1.30	2.53
alcohol		0.03			0.30	0.42

© Yossi Levy 2021

www.sci-princess.info

שאלה 3, סעיף ג

ג. מהו האמדן למקדם הרגרסיה b?

$\hat{b} = (\bar{y} - \hat{a})/\bar{x}$ $\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$

	Coefficients
Intercept	1.916
alcohol	

$\hat{a} = 1.916$

נתון כי ממוצע דירוגי היינות הוא 5.61 וכי ממוצע רמת האלכוהול של היינות שבמדגם הוא 10.16.

$\bar{y} = 5.61$ $\bar{x} = 10.16$

$\hat{b} = (5.61 - 1.916)/10.16 = 0.364$

שאלה 3, סעיף ד

ד. האם תדחו את ההשערות כי מקדמי הרגרסיה a ו-b שונים מאפס?

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1.916	0.31			1.30	2.53
alcohol		0.03			0.30	0.42

אין p-value אבל יש רווחי סמך

רווחי הסמך לא מכילים את 0 ולכן דוחים את ההשערות

© Yossi Levy 2021

www.sci-princess.info

שאלה 3, סעיפים ה-ו

ה. מהו אחוז השונות המוסברת על ידי הרגרסיה?

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	81.84	81.84		
Residual	596	347.79	0.58		
Total	597	429.63			

$429.63 - 347.79 = 81.84$

$$R^2 = \frac{81.84}{426.63} = 0.190$$

ו. מהו מקדם המתאם בין איכות היין ורמת האלכוהול?

$$R = \sqrt{0.190} = 0.436$$

יש לזכור כי הסימן של מקדם המתאם זהה לסימן של $\hat{b}$. אם $\hat{b}$ שלילי אז גם מקדם המתאם חייב להיות שלילי

שאלה 3, סעיף ז

ז. הסבירו את ההשפעה של שינוי במשתנה המוסבר על הניבוי של ערך המשתנה המסביר.

משוואת הרגרסיה היא: $\text{wine quality} = 1.916 + 0.363 \cdot \text{alcohol}$

על פי המודל, עליה של אחוז אחד ברמת האלכוהול תגדיל את התחזית לאיכות היין ב-0.363 נקודות

יש לשים לב כי העלאת אחוז האלכוהול לא בהכרח תגרום לשיפור באיכות היין!
מודל הרגרסיה אינו מראה קשר סיבתי אלא מתאם בלבד.
מצד שני, המודל לא שולל קיום של קשר סיבתי, וייתכן כי קשר כזה אכן קיים

© Yossi Levy 2021
www.sci-princess.info