תרגיל 4

שאלה 1

המודל :

Y - מספר הפטירות מסרטן האף

Age - גיל (קטגוריאלי - אורדינלי)

Year - מועד תחילת העסקה (קטגוריאלי - אורדינלי)

Exposure - מספר שנות החשיפה (קטגוריאלי - אורדינלי)

T - שנות אדם ביחידות של 100,000

Y(Age,Year,Exposure)~Pois(T(Age,Year,Exposure)xρ(Age,Year,Exposure))

ρ = שיעור

X – 0 = אינו שייך לקבוצה, 1 = שייך לקבוצה

$$\ln(T \times \rho) = \ln(T) + \ln(\rho) = \ln(T) + \beta_0 + \beta_{Age=2} \times X_{Age=2} + \beta_{Age=3} \times X_{Age=3} +$$
$$+ \beta_{Age=4} \times X_{Age=4} + \beta_{Year=2} \times X_{Year=2} + \beta_{Year=3} \times X_{Year=3} + \beta_{Year=4} \times X_{Year=4} +$$
$$+ \beta_{Exposure=2} \times X_{Exposure=2} + \beta_{Exposure=3} \times X_{Exposure=3} + \beta_{Exposure=4} \times X_{Exposure=4} +$$
$$+ \beta_{Exposure=5} \times X_{Exposure=5}$$

מניתוח הנתונים (טבלה בעמוד הבא) ניתן לראות כי חשיפה לניקל מהווה גורם סיכון לתמותה מסרטן האף עם קשר מנה – תגובה בין משך החשיפה לבין מוות מסרטן האף.

הסיכון היחסי לתמותה בקרב אלו שנחשפו 0.5 עד 4 שנים לעומת אלו שלא נחשפו היה 4.94 (לא מובהק, p=0.127).

הסיכון היחסי לתמותה בקרב אלו שנחשפו 4.5 עד 8 שנים לעומת אלו שלא נחשפו היה 5.76 (לא מובהק, p=0.097).

הסיכון היחסי לתמותה בקרב אלו שנחשפו 8.5 עד 12 שנים לעומת אלו שלא נחשפו היה 10.54 (מובהק, p=0.028).

הסיכון היחסי לתמותה בקרב אלו שנחשפו 12.5 שנים או יותר לעומת אלו שלא נחשפו היה 16.73 (מובהק, p=0.012). רווח הסמך מ- 1.9 ל- 150, בהתאם לפלט.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-4.666	1.3189	-7.251	-2.081	12.518	1	.000	.009	.001	.125
[Age=4]	3.428	.7816	1.896	4.960	19.234	1	.000	30.810	6.659	142.559
[Age=3]	2.482	.7591	.994	3.969	10.688	1	.001	11.961	2.702	52.954
[Age=2]	1.673	.7521	.198	3.147	4.946	1	.026	5.326	1.220	23.262
[Age=1]	0[a]	.	.	.	.	.	.	1	.	.
[Year=4]	-1.126	.4529	-2.014	-.238	6.181	1	.013	.324	.133	.788
[Year=3]	.054	.4681	-.864	.971	.013	1	.909	1.055	.422	2.641
[Year=2]	.619	.3713	-.109	1.347	2.779	1	.095	1.857	.897	3.845
[Year=1]	0[a]	.	.	.	.	.	.	1	.	.
[Exposure=5]	2.817	1.1183	.626	5.009	6.348	1	.012	16.734	1.869	149.791
[Exposure=4]	2.355	1.0701	.257	4.452	4.842	1	.028	10.536	1.294	85.809
[Exposure=3]	1.751	1.0552	-.317	3.819	2.754	1	.097	5.761	.728	45.574
[Exposure=2]	1.598	1.0475	-.455	3.651	2.327	1	.127	4.943	.634	38.512
[Exposure=1]	0[a]	.	.	.	.	.	.	1	.	.
(Scale)	1[b]									

Dependent Variable: Deaths

Model: (Intercept), Age, Year, Exposure, offset = LNPY100000

a. Set to zero because this parameter is redundant.b. Fixed at the displayed value.

על מנת לחשב את שיעורי התמותה נשתמש במודל :

$$\ln(\rho) = \beta_0 + \beta_{Age=2} \times X_{Age=2} + \beta_{Age=3} \times X_{Age=3} +$$
$$+ \beta_{Age=4} \times X_{Age=4} + \beta_{Year=2} \times X_{Year=2} + \beta_{Year=3} \times X_{Year=3} + \beta_{Year=4} \times X_{Year=4} +$$
$$+ \beta_{Exposure=2} \times X_{Exposure=2} + \beta_{Exposure=3} \times X_{Exposure=3} + \beta_{Exposure=4} \times X_{Exposure=4} +$$
$$+ \beta_{Exposure=5} \times X_{Exposure=5}$$

נציב את ה-β שקיבלנו בטבלה בעמוד הקודם ונכניס 0 ו-1 בערכי ה-X בהתאמה.

$$\ln(\rho) = \beta_0 + \beta_{Age=3} + \beta_{Year=2} + \beta_{Exposure=2} = -4.666 + 2.482 + 0.619 + 1.598 = 0.033$$

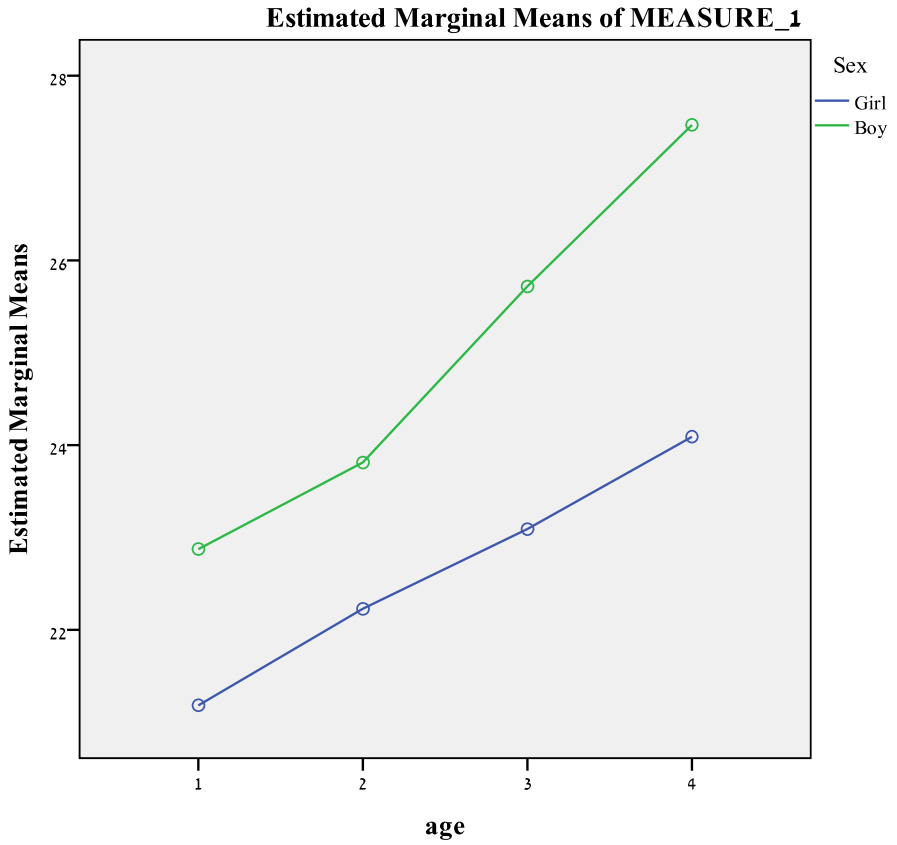
כעת נהפוך ליחידות של שיעור

$$e^{0.033} = 1.033$$

מכיוון שמראש ביטאנו את T ביחידות של 100,000 שנות אדם השיעור שהתקבל הוא שיעור הנאמד לתמותה מסרטן האף ל-100,000 שנות אדם בקרב אלו שבגיל 17.5 עד 34.9, התחילו לעבוד בין 1910 ל-1914 והייתה להם בין 0.5 ל-4 שנות חשיפה.

שאלה 2

סעיף 1 :



סעיף 2 :

מתוך בחינת השונויות ניתן לראות כי ניתן להשתמש במבחן Sphericity (p=0.2).

ישנו הבדל מובהק בין הגילאים (p<0.001).

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
age	Sphericity Assumed	209.437	3	69.812	35.347	.000
	Greenhouse-Geisser	209.437	2.602	80.503	35.347	.000
	Huynh-Feldt	209.437	3.000	69.812	35.347	.000
	Lower-bound	209.437	1.000	209.437	35.347	.000

מתוך הטבלה מטה ניתן לראות כי מרבית הקשר הינו לינארי (מובהק - $p<0.001$) ומיעוטו ריבועי (אינו מובהק - $p=0.347$) או שלישוני (אינו מובהק - $p=0.774$).

Tests of Within-Subjects Contrasts

Measure:MEASURE_1

Source	age	Type III Sum of Squares	df	Mean Square	F	Sig.
age	Linear	208.266	1	208.266	87.999	.000
	Quadratic	.959	1	.959	.920	.347
	Cubic	.212	1	.212	.084	.774

<u>סעיף 3</u> :

קיים הבדל מובהק בין בנים לבנות ($p=0.005$).

Tests of Between-Subjects Effects

Measure:MEASURE_1

Transformed Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	59118.502	1	59118.502	3910.836	.000
Sex	140.465	1	140.465	9.292	.005
Error	377.915	25	15.117		

<u>סעיף 4</u> :

סיכום גלובלי מצביע על כך שאין אינטראקציה מובהקת ($p=0.078$).

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
age * Sex	Sphericity Assumed	13.993	3	4.664	2.362	.078
	Greenhouse-Geisser	13.993	2.602	5.378	2.362	.088
	Huynh-Feldt	13.993	3.000	4.664	2.362	.078
	Lower-bound	13.993	1.000	13.993	2.362	.137

אמנם, כשמפרטים את האנטראקציה למרכיבים (בדומה לפירוק של השפעת הגיל), רואים שיש הבדל מובהק ($0.033=p$) בין בנים ובנות במרכיב הלינארי, אך לא במרכיבים האחרים:

Tests of Within-Subjects Contrasts

Measure:MEASURE_1

Source	age	Type III Sum of Squares	df	Mean Square	F	Sig.
age * Sex	Linear	12.114	1	12.114	5.119	0.033
	Quadratic	1.2	1	1.2	1.152	0.293
	Cubic	0.679	1	0.679	0.27	0.608
Error(age)	Linear	59.167	25	2.367		
	Quadratic	26.041	25	1.042		
	Cubic	62.919	25	2.517		

עיון בנתונים מראה שקצב הגידול הלינארי הוא מעט גדול יותר, ובאופן מובהק, בקרב הבנים.

<u>סעיף 5</u> :

האם ישנו הבדל בין הגילאים – המבחן בתוך הנבדקים.

האם ישנו הבדל בין בנים לבנות – המבחן בין נבדקים.

האם ישנה אינטראקציה בין גיל ומין – המבחן בתוך הנבדקים.

<u>שאלה 3</u>

<u>סעיף 1</u> :

ברמה 1, בתוך הנבדקים, המודל הוא: $y_{i,j} = \beta_{0,i} + \beta_{1,i} t_{i,j} + \beta_{2,i} t_{i,j}^2 + \varepsilon_{i,j}$ כאשר $y_{i,j}$

התצפית ה- j על נבדק i ומתקבל בגיל $t_{i,j}$.

ברמה 2, בין נבדקים:

$$\beta_{0,i} = \beta_0 + \beta_{0,Sex} Sex_i + \delta_{0,i}$$

$$\beta_{1,i} = \beta_1 + \beta_{1,Sex} Sex_i + \delta_{1,i}$$

$$\beta_{2,i} = \beta_2$$

לשים לב שהמודל כולל, לכל נבדק, מרכיב ריבועי (רמה 1) אבל מניחים שמרכיב זה זהה לכל הנבדקים (רמה 2), כי אין לו אינטראקציה עם מין ואין לו מרכיב אקראי.

<u>סעיפים 2 עד 5:</u>

אציג קודם פלט למודל הנ"ל. כפי שתראו, המרכיב הריבועי אינו מובהק. לכן רצוי גם להתאים מודל בלעדיו ומסקנות סופיות מן הניתוח אסיק מהמודל ללא מרכיב ריבועי.

להלן תוצאות מהמודל הכולל מרכיב ריבועי:

Estimates of Fixed Effects[b]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	19.697106	3.930839	74.676	5.011	.000	11.865916	27.528297
[Sex=0]	1.032102	1.475955	66.496	.699	.487	-1.914326	3.978530
[Sex=1]	0[a]	0	.	.	.	.	.
Age	.147801	.728135	73.710	.203	.840	-1.303134	1.598736
AgeSq	.028935	.032899	73.440	.880	.382	-.036626	.094496
[Sex=0] * Age	-.304830	.124622	66.496	-2.446	.017	-.553611	-.056049
[Sex=1] * Age	0[a]	0	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Dependent Variable: Dist.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Repeated Measures	Variance	1.870286	.308643	6.060	.000	1.353439	2.584504
Intercept [subject = Subject]	Variance	2.417413	1.467879	1.647	.100	.735347	7.947113
Age [subject = Subject]	Variance	.007722	.010547	.732	.464	.000531	.112271

a. Dependent Variable: Dist.

לשים לב שהגורם המייצג את "מין" הוא 0 לבנים ו- 1 לבנות. במודל הראשון, המשואה הנאמדת לבנים היא: $19.70 + 0.15t + 0.029t^2$ כאשר t מסמן גיל. המשואה לבנות היא:
$20.73 - 0.16t + 0.029t^2$. שני המקדמים הקושרים את המרחק לגיל אינם מובהקים (מובהקות 0.84 לגורם הלניארי ו- 0.38 לריבועי). אחת הסיבות לכך היא שיש ביניהם מתאם די גדול. לכן הגיוני להוציא את המרכיב הריבועי ושוב להתאים את המודל.

להלן התוצאות מן המודל ללא הגורם הריבועי:

Estimates of Fixed Effects[b]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	16.340625	.940869	67.094	17.368	.000	14.462691	18.218559
[Sex=0]	1.032102	1.474059	67.094	.700	.486	-1.910055	3.974260
[Sex=1]	0[a]	0	.	.	.	.	.
Age	.784375	.079442	67.094	9.874	.000	.625812	.942938
[Sex=0] * Age	-.304830	.124462	67.094	-2.449	.017	-.553250	-.056409
[Sex=1] * Age	0[a]	0	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Dependent Variable: Dist.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Repeated Measures	Variance	1.864595	.305688	6.100	.000	1.352183	2.571186
Intercept [subject = Subject]	Variance	2.416803	1.464855	1.650	.099	.736745	7.928029
Age [subject = Subject]	Variance	.007747	.010524	.736	.462	.000540	.111041

a. Dependent Variable: Dist.

עכשיו המרכיב הלניארי מובהק מאוד (0.000) וכמו-כן האנטראקציה בין גיל ומין (0.017). מקדם האנטראקציה שלילי וזה מצביע על כך שהשיפוע (כלומר קצב הגידול) אצל הבנות נמוך מהשיפוע בקרב הבנים. המשואה הנאמדת לבנים היא: $16.34 + 0.78t$ כאשר t מסמן גיל. המשואה לבנות היא: $17.37 + 0.48t$. לשים לב סילוק הגורם הריבועי נותן שיפוע חיובי אצל הבנות, לעומת השיפוע השלילי במודל הראשון. ערכו של מקדם (וגם הסימן שלו) תלויים מאוד בגורמים הנוספים הנמצאים במשוואה. במודל הראשון, הגורם הריבועי תופס תפקיד "דומיננטי" בהבעת הקשר בין מרחק לבין גיל ומשבש את היכולת לראות את הקשר, שהוא בעיקר לניארי. ביישומים רבים, רצוי לבדוק האם גורם ריבועי נחוץ. כאשר (כמו כאן) מגלים שהגורם אינו נחוץ, עדיף להתאים מודל חדש בלעדיו שיהיה קל יותר לפרש.

נסתכל על הגורמים הקשורים למין. ההשפעה הראשית של מין היא 1.03 ומסמן שעל פי הנתונים שלנו, מעריכים שהמרחק הוא גדול יותר אצל בנות, ב- 1.03 ס"מ, בגיל 0 (כלומר בלידה). אבל יש טעות תקן גדולה (1.47) והמקדם אינו מובהק – אין פלא שנתונים מגיל 8 עד גיל 14 לא נותנים הרבה מידע על מידת השוני בגיל 0 (אפילו בהנחה הדי מפוקפקת שהמודל ניתן לאקסטרפולציה עד גיל 0). האנטראקציה עם השיפוע על גיל כן מובהקת (0.017) ולכן יש עדות משכנעת שקצב הגידול של בנים שונה מהקצב של בנות.

גורם השונות הקשור לרמה 1 (פיזור סביב העקומה הלניארית בקרב כל ילד וילד) נאמד על 1.86. נזכור שמדובר בשונות. סטיית התקן (שורש השונות) היא 1.36. המרכיב הבא משקף את השוני, בין נבדק לנבדק, בחותכים שלהם (ז"א בגובה הקו שלהם כשמותחים את הקו מאזור הנתונים אחורנית לגיל 0). חלק מן השוני בחותכים עשוי להיות כתוצאה מהבדל בין בנים ובנות, אמנם ראינו כבר שאין עדות משכנעת לכך בנתונים. הפיזור הנותר בחותכים נאמד עם שונות של 2.42, כלומר עם סטיית תקן של 1.56. אנו רואים שהפיזור בין נבדק לנבדק מעל גיל 0 עוד יותר גדול מאשר הפיזור של הנתונים סביב הקווים. (פירוש הדבר: אילו הסתכלתם על תרשים פיזור של כל המרחקים מול גיל ושאלתם: למה יש כל כך הרבה פיזור? חלק חשוב של הפיזור נובע משינויים בין נבדק לנבדק, ולא רק לפיזור בתוך נבדק, לאורך סדרת הנתונים שלו.) למרכיב הקשור לשיפוע שונות של 0.0077, כלומר סטיית תקן של 0.088. כדי להבין מספר זה, חשוב לראות מהם השיפועים הטיפוסיים. בקרב בנים, שיפוע טיפוסי הוא 0.78 ובקרב בנות 0.48. עד כמה יש שוני בשיפוע בין ילד אחד לילד אחר? בהתאם לסטיית תקן של 0.088. למשל, אם נחשוב על ה"כלל" של התפלגות נורמלית בו כ- 95% מן הנתונים הם עד 2 סטיות תקן מן הממוצע, נראה של- 95% מן הבנים שיפוע שבין 0.60 ו- 0.96. אנו רואים ששני המרכיבים האקראיים שנכנסו למודל ברמה 2 אינם משיגים מובהקות סטטיסטית. עצתי כאן היא, בכל זאת, להשאיר אותם במודל. התפקיד החשוב שלהם הוא לבטא את האפשרות לתלות בתוך נבדק ולדאוג שנעריך בהתאם את המובהקות של מקדמי הרגרסיה (שהם המטרה העיקרית של ההסקה הסטטיסטית כאן).

שאלה 4

התשובות כאן דומות לאלו של שאלה 3. כמובן שהוצאת חלק מן הנתונים משנים את הערכים של המקדמים הנאמדים ומרכיבי השונות הנאמדים. אציג כאן רק את הפלט של המודל ללא המרכיב הריבועי.

Estimates of Fixed Effects[b]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	16.351734	.976588	66.588	16.744	.000	14.402235	18.301234
[Sex=0]	1.376917	1.550009	66.748	.888	.378	-1.717129	4.470962
[Sex=1]	0[a]	0	.	.	.	.	.
Age	.785567	.083596	63.230	9.397	.000	.618525	.952608
[Sex=0] * Age	-.333796	.131876	64.260	-2.531	.014	-.597229	-.070364
[Sex=1] * Age	0[a]	0	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Dependent Variable: Dist.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Repeated Measures	Variance	1.911975	.325747	5.870	.000	1.369182	2.669951
Intercept [subject = Subject]	Variance	2.482398	1.506853	1.647	.099	.755404	8.157624
Age [subject = Subject]	Variance	.007305	.011006	.664	.507	.000381	.139991

a. Dependent Variable: Dist.