BIOS 6611 Homework 9 *Answer Key*

Due Monday, November 26, 2018 by midnight to Canvas Assignment Basket

The data for this exercise (lead2.sas7bdat SAS dataset available on Canvas in data repository and on assignment page) are from a study of the neurological and psychological effects of environmental lead exposure in a population of children who had lived near a lead smelter in El Paso, Texas. The study was conducted because previous work suggested that blood-lead levels of 40-80 μg/100ml adversely affected the nervous system, and this level of absorption might be associated with behavior abnormalities. Thus, the study purpose was to look for a relationship between lead exposure and neurological damage or behavior problems in children.

The data set for this exercise comes from a group of 124 children who had participated in the survey of blood lead levels in the larger study and who had lived near the lead smelter for at least 12 of the 24 months preceding the study. Forty-six of these children (the "exposed group") had high blood-lead levels (≥ 40μg/100ml). The other children (the "unexposed group") had blood-lead levels less than 40μg/100ml. The study also collected data on the distance each child lived from the smelter, how long the child lived at the residence, and whether the child lived in that residence during the first 2 years of life.

After obtaining parental permission, children were evaluated individually by examiners who did not know the child's blood-lead level or exposure group. Children first underwent complete medical and neurological evaluations. All children were given a battery of neurological tests including the Wechsler intelligence scale for children (WISC) or the Wechsler preschool and primary scale of intelligence (WPPSI) used to measure intelligence (IQ).

id	Subject ID number
age	Age in years
sex	Sex (1 = male; 2 = female)
race	Race (1=African American; 0=Non-Hispanic White)
resdur	Years lived at current residence
miles	Distance of current residence from the smelter (miles)
first2y	Did the child live at current residence during the first 2 years of life? (1=yes; 0=no)
expose	Lead exposure group, blood lead levels >40μg/100ml (0 = not exposed; 1 = exposed)
iq	IQ test score

BIOS 6611: Assignment #9 ANSWER KEY

1) Use linear regression to examine the relationship between IQ (*iq*) and lead exposure (*expose*):

A) What is the unadjusted (crude) estimate for the association between IQ and lead exposure? Write a brief, but <u>complete</u>, summary of the relationship between IQ and lead exposure.

```
DATA lead;
      SET "~lead2.sas7bdat"; /* lead2 is the name of the SAS data set */
RUN;

CRUDE MODEL
PROC REG DATA=lead;
      MODEL iq = expose / CLB;
RUN;
```

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	Intercept	1	102.70513	1.76710	58.12	<.0001	99.20698	106.20328
expose	Lead exposure group	1	-7.77035	2.90130	-2.68	0.0084	-13.51376	-2.02693

There is a significant relationship between blood lead levels and IQ (p = 0.0084). On average, IQ is 7.77 points lower (95% CI: 2.03 to 13.52 points lower) in children with blood lead levels >40μg/100ml compared to children with blood lead levels ≤ 40μg/100ml.

B) Adjusting for the effect of race, what is the adjusted estimate for the association between IQ and lead exposure? Write a brief, but <u>complete</u>, summary of the relationship between IQ and lead exposure adjusting for race.

ADJUSTED MODEL

```
PROC REG DATA=lead;
      MODEL iq = expose race/ CLB;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	102.73167	1.79692	57.17	<.0001	99.17420	106.28914
expose	Lead exposure group	1	-7.47285	4.31551	-1.73	0.0859	-16.01655	1.07085
race	Race	1	-0.41404	4.43127	-0.09	0.9257	-9.18690	8.35882

Adjusting for race, the relationship between blood lead levels and IQ is not significant (p = 0.0859). Within each race group, on average, IQ is 7.47 points lower (95% CI: 16.02 point lower to 1.07 point higher) in children with blood lead levels >40μg/100ml compared to children with blood lead levels ≤ 40μg/100ml.

C) Is race a confounder of the association between IQ and lead exposure? Should you report the results from (A) or (B)? Justify your answer.

Using the operational definition, race is not a confounder. The parameter estimate for exposure barely changes (it changes by 3.8% or 4.0% depending on approach used) when adjusting for race:

To Calculate %Change: $\frac{7.77035-7.47285}{7.77035} \times 100 = 3.8\%$ (method favored by biostatisticians)

To Calculate %Change: $\frac{7.77035-7.47285}{7.47285} \times 100 = 4.0\%$ (method favored by epidemiologists)

Using the classical definition, race is not a confounder. Although there is a very strong association between exposure and race (from the PROC FREQ output below we see that 78.26% of the exposed are African-American, versus only 6.4% of the unexposed are African American), meeting the first criterion for confounding; but race is not associated with IQ independent of its association with exposure (race is not significant in the adjusted model).

The results from (A) should be reported since race is not a confounder. Note that exposure is no longer significant in the adjusted model, but this is due to a dramatic increase in its standard error which occurred because the model adjusts for a variable associated with exposure but not with disease (recall the relative precision equation from lecture).

If you fit the covariate model (output on next page), you can calculate the "change in estimate" directly:

$$\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}} = \hat{\gamma}_X \times \hat{\beta}_Z$$

$$-7.77035 - (-7.47285) = 0.71851 \times -0.41404$$

$$-0.2975 = -0.2975$$

Commented [KAM1]: 30 points total:

10 points for assessing confounding

5 points for correct % change calculation

15 points for general summary (including reporting the results from A)

```
PROC FREQ DATA=lead;
      TABLES race*expose/ CHISQ;
RUN;
```

COVARIATE MODEL

```
PROC REG DATA=lead;
      MODEL race = expose/ CLB;
RUN;
```

Table of race by expose

race(Race)	expose(Lead exposure group)		
Frequency Percent Row Pct Col Pct	0	1	Total
0	73 58.87 87.95 93.59	10 8.06 12.05 21.74	83 66.94
1	5 4.03 12.20 **6.41**	36 29.03 87.80 **78.26**	41 33.06
Total	78 62.90	46 37.10	124 100.00

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	0.06410	0.03625	1.77	0.0795	-0.00766	0.13587
expose	Lead exposure group	1	**0.71851**	0.05952	12.07	<.0001	0.60068	0.83633

2) Use linear regression to examine the relationship between IQ (*iq*) and the distance of the current residence from the smelter (*miles*). In this question you will examine whether the magnitude of the association between IQ (the response) and distance of the residence from the smelter (the primary explanatory variable) depends on whether the child was exposed during the first two years of life (i.e., if they lived in the current residence during the first two years of life).

A) Write down the regression equation for the regression of *IQ* on *miles*, *first2y*, and the interaction between *miles* and *first2y*. Provide an interpretation for each of the coefficients in the model (including the intercept).

```
/* New DATA step to create new variables */
DATA lead2;
      SET lead;
      IF first2y = 1 THEN notfirst2y = 0;
      IF first2y = 0 THEN notfirst2y = 1;

      milesf2y = miles*first2y;
      milesNf2y = miles*notfirst2y;

      LABEL notfirst2y = 'Not exposed first 2years'
            milesf2y = 'miles*first2y'
            milesNf2y = 'miles*notfirst2y';
RUN;

/* NOT exposed during first 2 years as reference group*/
PROC REG DATA=lead2;
      MODEL iq = miles first2y milesf2y / CLB;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	98.45133	4.38619	22.45	<.0001	89.76699	107.13568
miles	Distance from smelter (miles)	1	0.59134	2.10468	0.28	0.7792	-3.57577	4.75845
first2y	Exposed during first 2yrs	1	-19.61160	7.85829	-2.50	0.0139	-35.17046	-4.05275
milesf2y	miles*first2y	1	17.65461	5.48109	3.22	0.0016	6.80244	28.50678

IQ = 98.451 + 0.591×miles + (-19.611)×first2y + 17.655×miles×first2y

Intercept = 98.451: The expected IQ for children living 0 miles from the smelter who did not live in the residence during the first two years of life is 98.45 points.

Miles=0.591: *For children who did not live in the current residence during the first two years of life*, IQ increases, on average, by 0.59 points for every mile of distance the child currently lives from the smelter.

First2y=-19.611: *For children who live 0 miles from the smelter*, IQ scores, on average, are 19.611 points lower for children who lived in the current residence during the first two years of life.

Miles*first2y = 17.65: This is the difference between the effect of miles for those exposed during the first 2 years of life compared to those not exposed during the first 2 years of life. For children exposed in the first two years of life, a one mile increase in distance from the smelter results in an IQ score that is 17.65 points higher, on average.

B) Test whether the relationship between *IQ* and *miles* depends on whether the child lived in the residence during the first two years of life.

The relationship between IQ and miles is significantly different for children who lived in the residence during the first two years of life compared to children who did not live in the residence during the first two years of life (p = 0.0016).

C) What is the regression equation for children who lived in the current residence during the first two years of life? For those who didn't live in the residence during the first two years of life?

You could use the original model and calculate the intercept and slope by hand for the children who lived in the current residence during the first 2 years of life, or recode your model to get this information directly from the SAS output (see output below for use of BY statement in PROC REG or use of reverse coding).

Using output from 2A:

Regression equation for children who lived in the current residence during the first 2 years:

$$IQ = 78.83973 + 18.24595 \times miles$$

Regression equation for children who did not live in the residence during the first 2 years:

$$IQ = 98.45133 + 0.59134 \times miles$$

Recoding for two separate models or reverse coding:

```
/* Question 2c if you choose to recode and fit two models */
/* Two Models */
PROC SORT DATA=lead2;
      BY first2y; /* sort data by first2y classification */
RUN;

/* Calculate estimated intercept and miles coef for first2y = YES/1 */
PROC REG DATA=lead2;
      MODEL iq = miles  / CLB;
      BY first2y ; /* produces separate parameter est sorted by first2y */
RUN;

/* Reversed coding for exposed during first 2 years */
PROC REG DATA=lead2;
      MODEL iq = miles notfirst2y milesNf2y / CLB;
RUN;
```

OUTPUT FOR PROC REG WITH BY STATEMENT:

Parameter estimates for children who lived in current house during first 2 yrs:

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	Intercept	1	98.45133	4.56587	21.56	<.0001	89.37617	107.52650
miles	Distance from smelter (miles)	1	0.59134	2.19090	0.27	0.7879	-3.76330	4.94598

Parameter estimates for children who lived in current house during first 2 yrs:

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	Intercept	1	78.83973	5.75698	13.69	<.0001	67.12707	90.55239
miles	Distance from smelter (miles)	1	18.24595	4.46844	4.08	0.0003	9.15485	27.33705

OUTPUT FOR PROC REG WITH REVERSE CODING:

Parameter estimates with reverse coding for first2y:

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	Intercept	1	78.83973	6.52028	12.09	<.0001	65.93003	91.74943
miles	Distance from smelter (miles)	1	18.24595	5.06089	3.61	0.0005	8.22574	28.26617
notfirst2y	Not exposed first 2years	1	19.61160	7.85829	2.50	0.0139	4.05275	35.17046
milesNf2y	miles*notfirst2y	1	-17.65461	5.48109	-3.22	0.0016	-28.50678	-6.80244

D) Provide a brief, but <u>complete,</u> summary of the relationship between IQ and distance of the current residence from the smelter, accounting for any observed interaction with exposure during the first two years of life. For your summary, include a scatterplot of *IQ* versus *miles*, using different symbols and separate regression lines for children who lived in the residence during the first two years of life and for those who didn't live in the residence during the first two years of life. Be sure to comment on the graph in your summary.

The relationship between the distance a child lives from the smelter and IQ differs significantly according to whether or not the child lived in the residence during the first 2 years of life (p = 0.0016). There is not a significant association between distance a child lives from the smelter and IQ test scores for children who did not live in the residence during the first two years of life (p = 0.7792). For these children, IQ only increases an average of 0.591 points for every mile further the child lives from the smelter (95% CI: -3.576, 4.758). There is a significant association between distance a child lives from the smelter and IQ test scores among children who lived in the residence during the first 2 years of life (p = 0.0005). For these children, IQ increases an average of 18.246 points for every mile further the child lives from the smelter (95% CI: 8.226, 28.266). The difference in these relationships for children exposed during the first two years can be seen in the scatterplot with regression fits for each group, where those exposed have a positive linear relationship with increasing IQ as distance from smelter increases.

Note: last sentence 95% CI from reverse coding output in 2C.

```
/* Question 2d figures */
/* SGPlot Code for 2D */
PROC SGPLOT DATA=lead2;
      REG Y=iq X=miles / GROUP=first2y;
      FORMAT first2y;
RUN;

/* Plot for 2D */
PROC GPLOT DATA=lead2;
      PLOT iq*miles = first2y / VAXIS=axis1 HAXIS=axis2 LEGEND=legend1;
      SYMBOL1 I=rl VALUE=circle COLOR=black LINE=3 WIDTH=2;
      SYMBOL2 I=rl VALUE=dot COLOR=black LINE=1  WIDTH=2;
      AXIS1 LABEL = (FONT=ARIAL HEIGHT=2.5 ANGLE=90 POSITION=center  )
       VALUE=(FONT=ARIAL HEIGHT=2);
      AXIS2 LABEL = (FONT=ARIAL HEIGHT= 2.5 POSITION=center )
       VALUE=(FONT=ARIAL HEIGHT=2);
      LEGEND1 FRAME LABEL=(FONT=ARIAL HEIGHT= 1.5) VALUE=(FONT=ARIAL
HEIGHT=1.5)
         POSITION=(bottom inside right) ACROSS=1
         ;
      FORMAT first2y yesno.;
RUN;
```

Commented [KAM2]: 50 points total:

10 points for overall summary

10 points for scatterplot *with* regression lines

10 points for test of interaction and conclusion

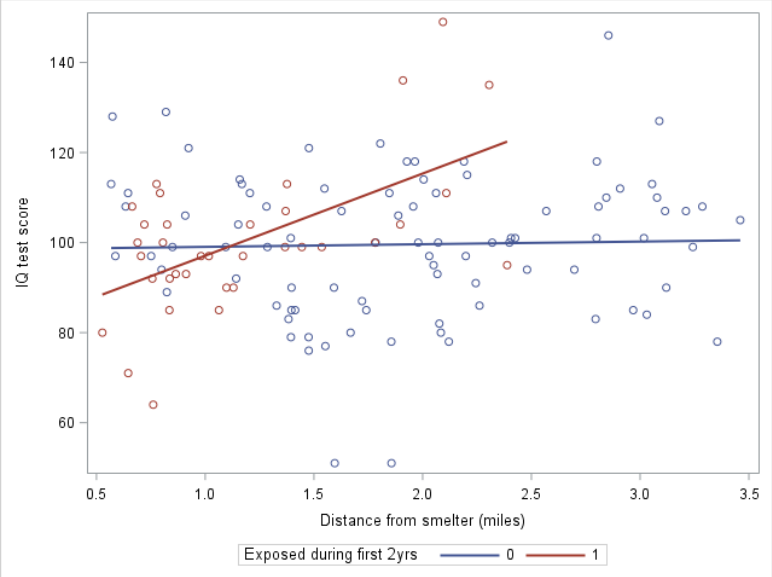
10 points for summary with CI for children living in residence

10 points for summary with CI for children not living in residence

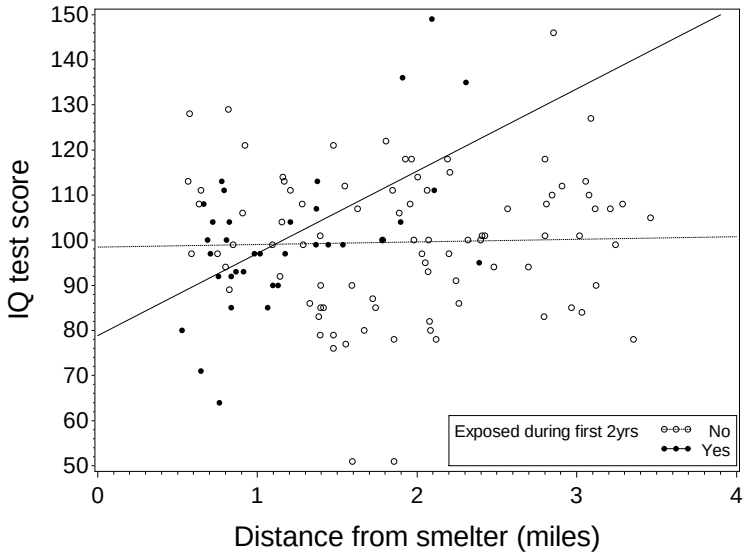
BIOS 6611: Assignment #9 ANSWER KEY

Figure 1. Relationship between IQ test scores and distance from the smelter, by exposure status during the first 2 years of life.

SGPLOT:



GPLOT:



3) By "hand", using the output from PROC REG, determine if the addition of both the covariate first2y and the interaction term miles*first2y in question 2A significantly contributes to the prediction of IQ, given the variable miles is included in the model 2A. Don't just use the commands in PROC REG to specify the partial F test, although you can check your answer with it. (i.e. what is the reduced and full model, the null hypothesis, and how do you test this?) [Hint: F2,120 = 3.07.]

Commented [KAM3]: 20 points total:

6 for the two correct models

6 for correct calculation based on models fit

8 for correct conclusion based on calculation

```
/* full model for 3 (from 2a) */
PROC REG DATA=lead2;
      MODEL iq = miles first2y milesf2y / CLB;
RUN;

/* reduced model for 3 */
PROC REG DATA=lead2;
      MODEL iq = miles / CLB;
RUN;
```

ANOVA table for model 2A with covariates miles, first2y and miles*first2y (full model):

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	**3106.32999**	1035.44333	4.38	0.0058
Error	120	28356	**236.29806**		
Corrected Total	123	31462			

ANOVA table for model with ONLY covariate miles (reduced model):

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	**443.85069**	443.85069	1.75	0.1889
Error	122	31018	254.24792		
Corrected Total	123	31462			

Using the partial F test we know our test statistic is:

$$F = \frac{[SS_{\text{model}}(full) - SS_{\text{model}}(reduced)]/k}{MS_{error}(full)} \sim F_{k,n-p-k-1}.$$

Using the output, we calculate:

$$F = \frac{(3106.32999 - 443.85069)/2}{236.29806} = 5.633731 \sim F_{2,124-1-2-1} = F_{2,120}, p = 0.004$$

With F=5.634 > $F_{2,120}$=3.07 (and p=0.004 < 0.05), we reject the null hypothesis and conclude that addition of the covariate first2y and the interaction term miles*first2y in Model 2A significantly contribute to the prediction of IQ, given the variable miles is included in the model 2A.