BIOS 6611 Homework 10 *Answer Key*

Due Monday, December 3, 2018 by midnight to Canvas Assignment Basket

The data for this assignment are from a study of demographic, dietary, and other behavioral determinants of plasma levels of retinol, beta-carotene, and other carotenoids. A summary of the study and a description of the variables in the *carotenoids.dat* data file are provided below:

Summary: Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. A cross-sectional study was designed to investigate the relationship between personal characteristics, dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. Study subjects (N = 315) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. The carotenoids.dat data file provides data for retinol and beta-carotene levels.

age	Age (years).
sex	Sex (1=Male, 2=Female).
smoke	Smoking status (1=Never; 2=Former; 3=Current Smoker).
bmi	Body Mass Index (weight/(height^2)).
vitamins	Vitamin Use (1=Yes, regularly; 2=Yes, irregularly; 3=No).
calories	Number of calories consumed per day.
fat	Fat consumed (grams per day).
fiber	Fiber consumed (grams per day).
alcohol	Number of alcoholic drinks consumed per week.
chol	Cholesterol consumed (mg per day).
betadiet	Dietary beta-carotene consumed (mg per day).
retdiet	Dietary retinol consumed (mg per day).
betaplas	Plasma beta-carotene (ng/ml).
retplas	Plasma Retinol (ng/ml).

1) The investigator hypothesizes that plasma beta-carotene levels may differ by smoking status. In this question, you will examine the relationship between plasma beta-carotene (the response) and smoking status (current smokers, former smokers, and never smokers).

A) Obtain the sample size, mean, standard deviation (SD), and standard error of the mean (SE) for plasma beta-carotene levels within each of the three smoking groups.

```
DATA carotenoids;
      INFILE "~\carotenoids.dat";
      INPUT age sex smoke bmi vitamins calories fat fiber alcohol chol
      betadiet retdiet betaplas retplas;
RUN;

/* create dummy variables */
data carotenoids;
      set carotenoids;

      *** Create dummy variables ****;
      IF smoke = 1 THEN smk_never = 1; ELSE smk_never = 0;
      IF smoke = 2 THEN smk_former = 1; ELSE smk_former = 0;
      IF smoke = 3 THEN smk_current = 1; ELSE smk_current = 0;

      *** Create variable for two groups with current status ****;
      IF smoke = 1 THEN group = 0;
      IF smoke = 2 THEN group = 1;
      IF smoke = 3 THEN group = 2;

      smk_nongrp = (group = 0 or group = 1); /* not currently smokers */
      smk_curgrp = (group = 2); /* current smokers */
run;

/* Question 1A */
PROC MEANS N MEAN STD STDERR data=carotenoids;
      VAR  betaplas;
      CLASS smoke;
      FORMAT smoke smoke.;
RUN;
```

Analysis Variable : betaplas					
smoke	N Obs	N	Mean	Std Dev	Std Error
Never	157	157	206.0509554	193.2085626	15.4197220
Former	115	115	193.4695652	191.6395246	17.8704778
Current	43	43	121.3255814	78.8116262	12.0186603

B) Fit a "reference cell" linear regression model (MODEL 1) regressing plasma beta-carotene levels, *betaplas* (the dependent variable) on smoking status (the independent variable). Make the never smokers the reference group. Write down the regression equation.

```
/* Questions 1B-1F */
PROC REG DATA=carotenoids;
      MODEL betaplas = smk_former smk_current/COVB CLB;
      TEST smk_former-smk_current;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	244625	122312	3.72	0.0254
Error	312	10271014	32920		
Corrected Total	314	10515638			

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	206.05096	14.48037	14.23	<.0001	177.55944	234.54247
smk_former	1	-12.58139	22.26974	-0.56	0.5725	-56.39924	31.23646
smk_current	1	-84.72537	31.22916	-2.71	0.0070	-146.17176	-23.27899

Covariance of Estimates			
Variable	**Intercept**	**smk_former**	**smk_current**
Intercept	209.6809913	-209.6809913	-209.6809913
smk_former	-209.6809913	495.94112724	209.6809913
smk_current	-209.6809913	209.6809913	975.26042464

Regression Equation:

$$E(\text{beta-carotene}) = \beta_0 + \beta_{former} I_{former} + \beta_{current} I_{current}$$

C) Using MODEL 1, is smoking status significantly associated with plasma beta-carotene levels? Write the null and alternative hypotheses in terms of the appropriate beta coefficient(s) and also in terms of the appropriate means, test the null hypothesis, and state your conclusion.

H_0: $\beta_{former} = \beta_{current} = 0$ H_0: $\mu_{never} = \mu_{former} = \mu_{current}$
H_A: $\beta_{former} \neq 0$ or $\beta_{current} \neq 0$ H_A: At least one μ different

$$F = \frac{MS(model)}{MS(error)} = \frac{122312}{32920} = 3.72 \sim F_{2,312}, p = 0.0254$$

There is a significant association between plasma beta-carotene levels and smoking status (p = 0.0254).

D) Using MODEL 1, do plasma beta-carotene levels differ between current smokers and never smokers? Write the null and alternative hypotheses in terms of the appropriate beta coefficient(s) and also in terms of the appropriate means, test the null hypothesis, and state your conclusion.

H_0: $\beta_{current} = 0$ H_0: $\mu_{never} = \mu_{current}$
H_A: $\beta_{current} \neq 0$ H_A: $\mu_{never} \neq \mu_{current}$

$$t = \frac{\hat{\beta}_{current}}{SE(\hat{\beta}_{current})} = \frac{-84.72537}{31.22916} = -2.71 \sim t_{312}, p = 0.0070$$

Reject the null hypothesis and conclude that never smokers have higher beta-carotene levels compared to current smokers.

E) Using MODEL 1, do plasma beta-carotene levels differ between former smokers and never smokers? Write the null and alternative hypotheses in terms of the appropriate beta coefficient(s) and also in terms of the appropriate means, test the null hypothesis, and state your conclusion.

H_0: $\beta_{former} = 0$ H_0: $\mu_{never} = \mu_{former}$
H_A: $\beta_{former} \neq 0$ H_A: $\mu_{never} \neq \mu_{former}$

$$t = \frac{\hat{\beta}_{former}}{SE(\hat{\beta}_{former})} = \frac{-12.58139}{22.26974} = -0.56 \sim t_{312}, p = 0.5727$$

Fail to reject the null hypothesis. There is not a significant difference in plasma beta-carotene levels between former and never smokers.

F) Using MODEL 1, do plasma beta-carotene levels differ between current smokers and former smokers? Write the null and alternative hypotheses in terms of the appropriate beta coefficient(s) and also in terms of the appropriate means, test the null hypothesis, and state your conclusion. (USE the variance-covariance matrix for the βs to answer this question).

Commented [KAM1]: 40 points total:

10 for code and output
5 for hypotheses using betas
5 for hypotheses using means
12 for construction and calculation of test by hand
8 for conclusion

H_0: $\beta_{former} - \beta_{current} = 0$ $\quad$ H_0: $\mu_{current} = \mu_{former}$

H_A: $\beta_{former} - \beta_{current} \neq 0$ $\quad$ H_A: $\mu_{current} \neq \mu_{former}$

$$t = \frac{\hat{\beta}_{former} - \hat{\beta}_{current}}{SE(\hat{\beta}_{former} - \hat{\beta}_{current})} = \frac{72.14398}{32.432076} = 2.224 \sim t_{312}, p = 0.02686$$

Where:

$$\hat{\beta}_{former} - \hat{\beta}_{current} = -12.58139 - (-84.72437) = 72.14398$$

$$SE(\hat{\beta}_{former} - \hat{\beta}_{current}) = \sqrt{Var(\hat{\beta}_{former}) + Var(\hat{\beta}_{current}) - 2Cov(\hat{\beta}_{former}, \hat{\beta}_{current})}$$
$$= \sqrt{495.94113 + 975.2604 - 2 \times 209.6810}$$
$$= 32.432076$$

Reject the null hypothesis. There is a significant difference in plasma beta-carotene levels between current and former smokers (p=0.02686).

Note, we can verify our hand calculation with the results of the TEST statement,

```
TEST smk_former-smk_current;
```

Test 1 Results for Dependent Variable betaplas				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	162896	4.95	0.0268
Denominator	312	32920		

2) For question 2 we will focus on fitting the "cell means" model instead of the "reference cell" model from question 1.

A) Fit a "cell means" linear regression model (MODEL 2) predicting plasma beta-carotene levels from smoking status. Write down the regression equation.

```
/* Question 2A-2C */
PROC REG DATA=carotenoids;
      MODEL betaplas = smk_never smk_former smk_current /NOINT COVB;
      TEST smk_never=smk_former=smk_current;                    /*2B*/
      TEST smk_never=smk_former,smk_never=smk_current;          /*Alternate 2B*/
      TEST smk_former-smk_current;                              /*2C */
      TEST .5*smk_never + .5*smk_former - smk_current;          /*2D*/
      TEST .5709*smk_never + .4291*smk_former - smk_current; /*2D, weighting by n*/
RUN;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11603208	3867736	117.49	<.0001
Error	312	10271014	32920		
Uncorrected Total	315	21874222			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
smk_never	1	206.05096	14.48037	14.23	<.0001
smk_former	1	193.46957	16.91922	11.43	<.0001
smk_current	1	121.32558	27.66911	4.38	<.0001

Covariance of Estimates			
Variable	smk_never	smk_former	smk_current
smk_never	209.6809913	0	0
smk_former	0	286.26013594	0
smk_current	0	0	765.57943334

Regression equation:

$$E(\text{beta-carotene}) = \beta_{never} I_{never} + \beta_{former} I_{former} + \beta_{current} I_{current}$$

B) Use the cell means model (MODEL 2) to test if smoking status is significantly associated with plasma beta-carotene levels.

Equivalent test statements:

```
TEST smk_never=smk_former=smk_current;
TEST smk_never=smk_former,smk_never=smk_current;
```

Test 1 Results for Dependent Variable betaplas				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	122312	3.72	0.0254
Denominator	312	32920		

Yes, plasma beta-carotene levels differ significantly between never, former, and current smokers (p = 0.0254).

C) Use the cell means model (MODEL 2) to test whether plasma beta-carotene levels differ between current smokers and former smokers. Write the null and alternative hypotheses in terms of the appropriate beta coefficient(s) and also in terms of the appropriate means, test the null hypothesis, and state your conclusion.

H_0: $\beta_{former} = \beta_{current}$ H_0: $\mu_{current} = \mu_{former}$

H_A: $\beta_{former} \neq \beta_{current}$ H_A: $\mu_{current} \neq \mu_{former}$

Test statement: `TEST smk_former-smk_current;`

Test 3 Results for Dependent Variable betaplas				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	162896	4.95	0.0268
Denominator	312	32920		

Reject the null hypothesis. There is a significant difference in plasma beta-carotene levels between current and former smokers (p = 0.0268).

Commented [KAM2]: 40 points total:

10 for code and output
5 for hypotheses using betas
5 for hypotheses using means
12 for test using SAS
8 for conclusion

D) Use the cell means model (MODEL 2) to test whether plasma beta-carotene levels differ between non-smokers (the average of never smokers and former smokers) and current smokers. Write the null and alternative hypotheses in terms of the appropriate beta coefficient(s) and also in terms of the appropriate means, test the null hypothesis, and state your conclusion.

H_0: $.5*\beta_{never} + .5*\beta_{former} = \beta_{current}$ H_0: $.5*(\mu_{never}+\mu_{former}) = \mu_{current}$

H_A: $.5*\beta_{never} + .5*\beta_{former} \neq \beta_{current}$ H_A: $.5*(\mu_{never}+\mu_{former}) \neq \mu_{current}$

Test statements:

```
TEST .5*smk_never + .5*smk_former - smk_current; /*Test 4: equal
weighting*/
TEST .5709*smk_never + .4291*smk_former - smk_current; /*Test 5:
weighting by n*/
```

Test 4 Results for Dependent Variable betaplas				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	227666	6.92	0.0090
Denominator	312	32920		

Test 5 Results for Dependent Variable betaplas				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	233645	7.10	0.0081
Denominator	312	32920		

The average beta-carotene levels for non-smokers (the average of never and current smokers) is significantly different than the average beta-carotene levels for current smokers (p = 0.0090).

3) Perform an independent samples t test comparing plasma beta-carotene levels in current smokers versus former smokers. Compare your results to those obtained in parts (1F and 2C) and explain any differences.

Commented [KAM3]: 20 points total

10 for t-test code and results
10 for comparison summary

```
/* Question 3 */
PROC TTEST DATA=carotenoids;
      CLASS smoke;
      VAR betaplas;
      WHERE smoke in (2,3); /* specify groups to compare */
      FORMAT smoke smoke.;
RUN;
```

smoke	N	Mean	Std Dev	Std Err	Minimum	Maximum
Former	115	193.5	191.6	17.8705	16.0000	1212.0
Current	43	121.3	78.8116	12.0187	25.0000	418.0
Diff (1-2)		72.1440	168.8	30.1819		

smoke	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Former		193.5	158.1	228.9	191.6	169.7	220.2
Current		121.3	97.0709	145.6	78.8116	64.9834	100.2
Diff (1-2)	Pooled	72.1440	12.5261	131.8	168.8	152.0	189.9
Diff (1-2)	Satterthwaite	72.1440	29.6010	114.7			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	156	2.39	0.0180
Satterthwaite	Unequal	154.6	3.35	0.0010

From 1F using regression model:

$$t = \frac{\hat{\beta}_{former} - \hat{\beta}_{current}}{SE\left(\hat{\beta}_{former} - \hat{\beta}_{current}\right)} = \frac{72.14398}{32.432076} = 2.224 \sim t_{312}, p = 0.02686$$

From t-test:

$$t = \frac{\left(\bar{Y}_{former} - \bar{Y}_{current}\right)}{SE\left(\bar{Y}_{former} - \bar{Y}_{current}\right)} = \frac{72.14398}{30.182} = 2.39 \sim t_{156}, p = 0.0180$$

The numerators of the two test statistics are identical. The denominators (the SE for the numerators) are different. The t test uses a pooled estimate of the SD based on only the two groups, while the regression model is using all three groups to estimate the pooled SD. Since the group left out (the never smokers) had the largest SD, the pooled SD for this t test is smaller than for the regression model.

Also, note that the degrees of freedom for the t statistic are larger for the regression model than the t test.