BIOS 6611 Homework 8 *Answer Key*

Due Monday, November 5, 2018 by midnight to Canvas Assignment Basket

Submit your ***complete*** SAS code as an appendix to your answers and include only relevant SAS output with your answers to each part.

1. From lectures 16-17, show that $SS_{Total} = SS_{Model} + SS_{Error}$ using the notes below.

Commented [KAM1]: 20 points

Consider $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. We can take material from lecture and rewrite it in terms of the residual ($\hat{e}_i$):

$$\frac{\partial}{\partial \beta_0} SS_{Error} = \frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right) = 0 \rightarrow \sum_{i=1}^{n} -2\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = -2\sum_{i=1}^{n}\hat{e}_i = 0$$

$$\frac{\partial}{\partial \beta_1} SS_{Error} = \frac{\partial}{\partial \beta_1}\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right) = 0 \rightarrow \sum_{i=1}^{n} -2X_i\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = -2\sum_{i=1}^{n} X_i\hat{e}_i = 0$$

Note 1: $\sum_{i=1}^{n}\hat{e}_i = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) = 0.$
Note 2: $\sum_{i=1}^{n} X_i\hat{e}_i = 0$
Hint:

$$SS_{Total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

From our lecture notes we have that $SS_{Error} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ and $SS_{Model} = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$. We therefore see that we need to show that $2\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)(\hat{Y}_i - \bar{Y}) = 0$ in order to show that $SS_{Total} = SS_{Model} + SS_{Error}$.

Based on our hints, we can first substitute the definition of $\hat{e}_i$:

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n}\hat{e}_i(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n}\hat{Y}_i\hat{e}_i - \bar{Y}\sum_{i=1}^{n}\hat{e}_i$$

Substituting in the definition of $\hat{Y}_i$:

$$\sum_{i=1}^{n}\hat{Y}_i\hat{e}_i - \bar{Y}\sum_{i=1}^{n}\hat{e}_i = \hat{\beta}_0\sum_{i=1}^{n}\hat{e}_i + \hat{\beta}_1\sum_{i=1}^{n}X_i\hat{e}_i - \bar{Y}\sum_{i=1}^{n}\hat{e}_i$$

Then we can apply notes 1 and 2:

$$\hat{\beta}_0\sum_{i=1}^{n}\hat{e}_i + \hat{\beta}_1\sum_{i=1}^{n}X_i\hat{e}_i - \bar{Y}\sum_{i=1}^{n}\hat{e}_i = \hat{\beta}_0(0) + \hat{\beta}_1(0) - \bar{Y}(0) = 0$$

Therefore, $SS_{Total} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + 0 = SS_{Model} + SS_{Error}$.

2. The following table presents the gender (0=Female; 1=Male), weight (kg), age (years), and plasma levels of total cholesterol (mg/100ml) for a sample of 7 patients suffering from hyperlipoproteinemia. The investigator is interested in examining the effect of weight on cholesterol levels.

Patient	*Gender*	*Cholesterol (mg/100ml)*	*Weight (kg)*	*Age (yr)*
1	0	254	57	23
2	1	402	79	57
3	0	288	63	28
4	1	354	84	46
5	0	220	30	34
6	1	451	76	57
7	0	405	65	52

For the data use the following:

```
DATA hw1;
INPUT id gender chol wtkg age;
chol2 = chol*chol;
wt2 = wtkg*wtkg;
cholwt = chol*wtkg;

LABEL wtkg = 'Weight (kg)'
        chol = 'Cholesterol (mg/100mL)'
        ;

DATALINES;
1  0    254   57    23
2  1    402   79    57
3  0    288   63    28
4  1    354   84    46
5  0    220   30    34
6  1    451   76    57
7  0    405   65    52
;
RUN;
```

A. Performing the regression of cholesterol (Y) on weight (X) in SAS, provide an ANOVA table and a parameter estimate table (only those tables, do NOT give all of the SAS output).

```
PROC REG;
      MODEL chol=wtkg/CLB;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	27382	27382	7.74	0.0388
Error	5	17699	3539.79126		
Corrected Total	6	45081			

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	Intercept	**1**	97.39533	89.78167	1.08	0.3275	-133.39580	328.18646
wtkg	Weight (kg)	**1**	3.72738	1.34017	2.78	0.0388	0.28236	7.17241

B. Write down the least-squares regression equation that describes the relationship between cholesterol (dependent variable) and weight (independent variable).

The regression equation is Cholesterol = 97.39533 + 3.72738 x Weight.

Note: Rounded values are also appropriate for the answer. Using $\hat{Y}$ (or Y if unsure how to create the y-hat notation, but be aware that it should be the predicted value of the outcome) and X in place of Cholesterol and Weight are also appropriate.

C. Inference about the intercept:

(I) What is the estimated intercept, and how would you interpret it?

The estimated intercept is 97.39533. It's interpretation is that the predicted cholesterol for an individual with a weight of 0 kg is 97.4 mg/100mL.

(II) Obtain a 95% confidence interval for the intercept and give its interpretation.

The 95% confidence interval for the intercept is (-133.4, 328.2). We are 95% confident that the true cholesterol for an individual with a weight of 0 kg is between -133.4 mg/100mL and 328.2 mg/100mL.

(III) Test the hypothesis that the true intercept is 0.

t = 1.08 and Pr > |t| = p = 0.3275. Since p>0.05, we fail to reject our null hypothesis that the true intercept is 0.

(IV) Is it scientifically interesting to test whether or not this intercept equals zero? Why or why not?

No, because a weight of 0 kg is biologically implausible (i.e., people have to weigh more than nothing). Additionally, even if the response wasn't biologically implausible, it falls outside the range of observed data and would result in extrapolation.

Note: Although we didn't reject the null hypothesis that the true intercept is 0 and it is not scientifically interesting, we should NOT remove the intercept from the model unless there is some scientific justification that the intercept should go through the origin (0,0).

D. Inference about the slope:

(I) What is the estimated slope, and how would you interpret it?

The estimated slope is 3.72738. For every one kilogram increase in weight, the cholesterol increases, on average, 3.727 mg/100mL.

(II) Obtain a 95% confidence interval for the slope and give its interpretation.

The 95% confidence interval for the slope is (0.282,7.172). We are 95% confident that cholesterol increases, on average, between 0.282 and 7.172 mg/100mL for a 1 kg increase in weight.

(III) Test the hypothesis that the true slope is zero.

t = 2.78, Pr > |t| = p =0.0388. Since p<0.05, we reject our null hypothesis that the true slope is 0.

E. Write a brief, but complete (i.e., include the point estimate, p-value, 95% CI, and summary/decision), summary of the effect of weight on cholesterol.

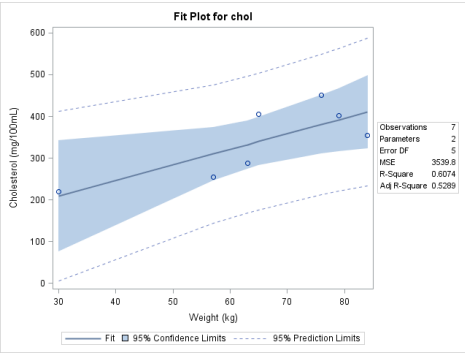
There is a significant association between weight and cholesterol level (p=0.0388). On average, cholesterol increases 3.73 mg/100mL per every 1 kilogram increase in weight (95% CI: 0.28, 7.17 mg/100mL).

Commented [KAM2]: 60 points:
15 for point estimate
15 for p-value
15 for 95% confidence interval
15 for summary/decision

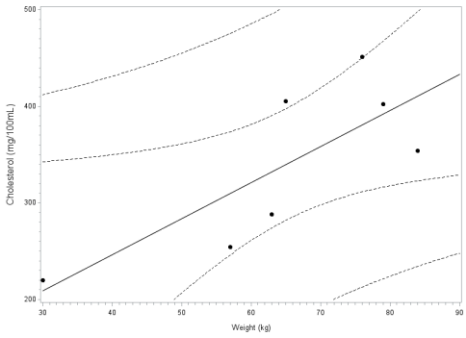
F. Use SAS to produce a scatterplot of cholesterol and weight along with the least squares regression line, the 95% confidence interval, and the 95% prediction interval. (Note: Depending on SAS version, PROC REG may also provide this type of plot.)

```
PROC GPLOT data=hw1;
   PLOT chol*wtkg=1 chol*wtkg=2 / OVERLAY VAXIS=axis1;
   SYMBOL1 INTERPOL=rlcli COLOR=black VALUE=dot;
   SYMBOL2 INTERPOL=rlclm COLOR=black VALUE=dot;
   AXIS1 LABEL = (FONT=ARIAL HEIGHT= 1.5 ANGLE=90 POSITION=center );
RUN;
```

Plot from PROC REG:



Plot from PROC GPLOT:



3. Read the paper by Moser and Stevens, “Homogeneity of Variance in the Two-Sample Means Test”, which is located on Canvas in the Paper Repository.

   In one paragraph, summarize the problem they studied, the methods they used, the results they obtained, their recommendations for statistical practice, and how you will apply the recommendations in the future.

   Moser and Stevens examine the behavior of the t-test under equal vs. unequal variances. They designed and executed a simulation study to determine the size (Type I error) and power of the test under various scenarios: assuming variances are equal, not equal (Smith/Welch/Satterthwaite – SWS test), and testing the equality of variances to decide which form of the t-test to use. When sample sizes are equal they conclude that the various strategies are very similar in terms of size and power. They also conclude that unless you have strong independent information that the variances are equal the best test to use is the unequal variances t-test as it preserves Type I error and power. Thus, in practice I would never test the equality of variances before deciding which test to use. Since I would rarely have strong information about the equality of variances I would always apply the SWS test.

Commented [KAM3]: 20 points

Answers can vary from what is written below as long as they are well written and justified.