BST762/STA632 And CPH 738 Homework Assignment #1

Due Thursday, January 24, 2019, at the beginning of class

1. Deal et al. [Journal of Applied Physiology, vol 46, pgs 467-475] measured ventilation volumes (1/min), which is our outcome of interest, from eight subjects under six different temperatures of inspired dry air. You have been given the dataset (ventilation_study). Temperature is in °Celsius, and can be thought of in the same way as time in longitudinal studies. For now, we will assume that the impact of temperature on mean volume is linear.

a) Fit the following linear regression model, ignoring the correlation among outcomes from the same subject:

$$Y_{ij} = \beta_0 + \beta_1 Temperature_{ij} + \epsilon_{ij}; \qquad j = 1, \ldots, 6; i = 1, \ldots, 8 \tag{1}$$

Here, Y_{ij} is the ventilation volume for subject i at temperature j. Specifically, $j = 1, 2, 3, 4, 5$, and 6 correspond to temperatures of -10, 25, 37, 50, 65, and 80, respectively.

i) What are $\hat{\beta}_1$, $\widehat{SE}(\hat{\beta}_1)$, and the degrees of freedom for the corresponding t-test? (5 points)

ii) Using a 5% significance level, do we have strong enough evidence to conclude that temperature is associated with volume? In other words, carry out the following test: $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$. (10 points)

b) An old approach to handling correlation in this type of situation is the derived variables approach. Because subjects are independent of each other, and outcomes within a subject are correlated, the derived variables approach will fit a separate linear regression model for each subject. The outcome from any given subject, i.e. the derived variable, will then be the estimated slope obtained from that subject's individual regression. Therefore, our derived dataset will consist of an estimated slope from each subject (8 observations). Our estimate for β_1 would then be the sample mean of these individual estimated slopes [$\hat{\beta}_1 = (1/8)\sum_{i=1}^{8} \hat{\beta}_{1i}$], and we can carry out a 1-sample t-test.

SAS code hints: In order to get results for each subject, use the line of code "by subject;" in proc reg. Then, create a new dataset with one variable that is the estimated slopes from each subjects' regression. Then use proc ttest to carry out the 1-sample t-test. Below is code you can use:

```
proc reg data=ventilation_study;
 model vent_volume = temperature;
 by subject;
run;

data slopes;
input estimate;
cards;
```

```
You need 8 lines of numbers (the estimated slopes) here

;
run;

proc ttest data=slopes;
 var estimate;
run;
```

i) What is $\hat{\beta}_1$, and what are $\widehat{SE}(\hat{\beta}_1)$ and the degrees of freedom for the 1-sample t-test? (5 points)

ii) Based on our derived dataset, and using a 5% significance level, do we have strong enough evidence to conclude that temperature is associated with volume? (5 points)

iii) Are results different from our analysis in which we ignored correlation? If so, then how? (10 points)

2. We now utilize a dataset (hgb.xls) that is described by Bush (2012) [Biostatistics: An Applied Introduction for the Public Health Practitioner, First Edition. Delmar, Cengage Learning.]. Pregnant women were followed from weeks 9 to 36, and we are interested in change in hemoglobin levels (g/dL). Hemoglobin level variables are hgb9 for week 9 and hgb36 for week 36. Women were categorized into the following: 1-Drank only tap water, 2-Drank only bottled water, or 3-Drank from both tap and bottled water (variable name is group). Another way tap water consumption was measured was continuously in liters (variable name is water). For now, ignore the other variables in the dataset. Use a 5% significance level for all tests.

a) Assume there were no drinking water categories. We are interested in whether or not hemoglobin levels changed from week 9 to week 36. Carry out a two-sample t-test, which ignores the correlation between these two measurements from the same woman, and a paired t-test (a 1-sample t-test on the variable "change"). To do a two-sample t-test, the following code may be helpful:

```
data twosample; set hgb; keep hgb9 hgb36; run;
data twosample; set twosample;
 y=hgb9; baseline=1; output;
 y=hgb36; baseline=0; output;
 drop hgb9 hgb36;
run;
proc sort data=twosample; by baseline; run;
```

i) What is the estimated mean change in hemoglobin using both methods, and what would you conclude from the two tests? (10 points)

ii) Which method appears to be more efficient. In other words, which method gave a smaller estimated standard error? (10 points)

b) Fit a linear regression model using hemoglobin change as the outcome of interest. Use the categorical water drinking variable to predict the outcome. In short, test if there are any mean differences in hemoglobin change between the three drinking groups. (Note that this is equivalent to a 1-Way ANOVA.) What is the R^2 from this model? (10 points)

c) Fit a linear regression model using hemoglobin change as the outcome of interest. This time, use the variable water, such that you fit a simple linear regression model. Test if there is an association between hemoglobin change and tap water consumption, interpret the estimated association and give the 95% CI for the slope/β_1, and give the R^2 from this model. How does this R^2 compare with R^2 from the model you fit in b, and therefore which model has better predictive accuracy?
Hint: To easily obtain the CI, you can use the following code (however, make sure you understand how to calculate the CI): (25 points)

```
proc glm data=hgb;
 model change = water / solution clparm;
run;
```

d) Fit the linear regression model from c), only this time include baseline hemoglobin (hemoglobin at week 9) as a covariate.

i) Given that tap water consumption is in the model, do we have strong enough statistical evidence to conclude that baseline hemoglobin level is associated with hemoglobin change? Explain. (5 points)

ii) Given that baseline hemoglobin level is in the model, do we have strong enough statistical evidence to conclude that tap water consumption is associated with hemoglobin change? Explain. (5 points)