Homework # 1
STOR 455
Due June 28, 2017

Show only relevant output (i.e. relevant tables and plots). Attach a copy of your SAS code. Restrict answers to one or two sentences.

Problem 1

Download the dataset from Sakai (ati.txt). The variables in this dataset are verbal scores (VERBAL), visual memory scores (VIS_MEM), and whether or not the participants received training (TRAINING, 0= no, 1 = yes). Read this dataset into SAS to answer the following questions:

a. Using SAS, construct and print out a grouped frequency table for VIS_MEM.
b. Write SAS code to create a new variable called VIS_MEM1 that is equal to VIS_MEM minus the mean of VIS_MEM, divided by the standard deviation of VIS_MEM. What is the mean and variance of the new variable? (Include SAS code for full credit, hint: type in the mean and standard deviation directly into the code)
c. Compute the means for VIS_MEM by TRAINING and output them. Then, plot the VIS_MEM means by TRAINING.
d. Generate a scatterplot of VIS_MEM versus VERBAL. Does there appear to be a relationship between the two variables?

Problem 2

The following sample data are IQ scores for 36 individuals:

134	131	126	132	138	106	136	85	121
120	135	106	113	145	126	124	130	144
129	114	123	121	116	129	91	101	146
146	119	132	132	125	123	130	131	115

Input the data into SAS and run the following analyses:

a. Run PROC UNIVARIATE on IQ. Report the measures of central tendency and explain what they tell us about the distribution of the IQ scores?
b. Generate a histogram, boxplot and a QQ normal plot of the data. Describe the shape of the distribution of IQ scores.

Problem 3

Assume a normally distributed population of resting heart rates with $\mu = 76$ and $\sigma = 5$. Use SAS to determine:

a. What is the probability of randomly selecting someone whose heart rate is below 53?

b. What is the probability of randomly selecting someone whose heart rate is either above 80, below 69, or between 73 and 77? Also, sketch a standard normal curve and shade the area representing the proportion.

Assignment #4
STOR 455
Due July 10, 2017

Show only relevant output (i.e. relevant tables and plots). Print a paper copy to turn in during class. You may want to attach a copy of your SAS code. Do not answer questions in SAS code.

From the book 3.6 a,b,c, 3.9, 3.18, 5.2, 5.5, 5.13

Using results from problems 5.5 and 5.13, find (using matrix methods):

1) Vector of estimated regression coefficients
2) Vector of predicted values
3) Hat matrix **H**
4) Vector of residuals

Assignment #5
STOR 455
Due July 13, 2017

Show only relevant output (i.e. relevant tables and plots). Print a paper copy to turn in during class. You may want to attach a copy of your SAS code. Do not answer questions in SAS code.

From the book 6.5 a,b, 6.6 a,b (this is the F test), 6.7 a, 6.8 a,b, 6.22 a,b,c,d,e

Assignment #6
STOR 455
Due July 17, 2017

Show only relevant output (i.e. relevant tables and plots). Print a paper copy to turn in during class. You may want to attach a copy of your SAS code. Do not answer questions in SAS code.

1. A random sample of 56 students who attended orientation prior to starting their first semester of college was taken. The variables are symptoms, daily hassles, and perceived level of social support. (symptoms.txt on Sakai)

    a. Using symptoms as the dependent variable, develop a regression model with daily hassles and perceived level of social support as the independent variables. Interpret the parameter estimates.
    b. Create an interaction model from the model in part (a). Be sure to center IVs. Let perceived level of social support be the moderator. Interpret parameter estimates (except the interaction term).

    For relevant parts of questions (c) through (g) you should use the online calculator found at http://www.quantpsy.org/interact/mlr2.htm

    c. Using the results from above, obtain regression equations for predicting symptoms for each of three levels of support: mean, 1 sd above the mean, and 1 sd below the mean. Write out the three regression equations and plot the 3 regression lines on a single plot.
    d. Report the values of the simple slopes for hassles at each support level as well as results of significance tests for each simple slope.
    e. Write a brief interpretation of the nature of the interaction effect, focusing on the plot from part (c) and the simple slopes from part (d).
    f. Obtain the region of significance for the simple slope for the effect of hassles on symptoms, treating support as the moderator. Provide the relevant table. Write a brief explanation of this result.
    g. Obtain the confidence band for the simple slope for the effect of hassles on symptoms, treating support as the moderator. Provide the relevant plot. Write a brief interpretation of this result.

2. A random sample of runners were assessed on the following variables (runners.txt on Sakai):

    RUNTIME = time to run 1.5 miles (in minutes)
    OXYGEN = oxygen use
    RUNPUL = heart rate while running
    AGE = Age in years
    RESTPUL = resting heart rate

WEIGHT = individual's weight
MALE: 0 = female, 1 = male
SMOKE: 2 = never smoked, 1 = used to smoke but quit
0 = current smoker
DRINK: Before running, drank: 3 = water, 2 = gatorade, 1 = powerade,
0 = caffeinated beverage

a. Using runtime as the dependent variable, develop a regression model with age and smoke as the independent variables. Interpret parameter estimates from the model.
b. Create an interaction model from the model in part (a). Let smoke be the moderator.
c. Create and interpret a plot to "probe" the interaction.
d. Create a regression of your own using the variables above (include one categorical variable). What did you find?

Assignment #7
STOR 455
Due July 19, 2017

Show only relevant output (i.e. relevant tables and plots). Print a paper copy to turn in during class. You may want to attach a copy of your SAS code. Do not answer questions in SAS code.

A random sample of runners were assessed on the following variables (runners.txt on Sakai):

RUNTIME = time to run 1.5 miles (in minutes)
OXYGEN = oxygen use
RUNPUL = heart rate while running
AGE = Age in years
RESTPUL = resting heart rate
WEIGHT = individual's weight
MALE: 0 = female, 1 = male
SMOKE: 2 = never smoked, 1 = used to smoke but quit
0 = current smoker
DRINK: Before running, drank: 3 = water, 2 = gatorade, 1 = powerade,
0 = caffeinated beverage

a. Using runtime as the dependent variable, develop a regression model with age and smoke as the independent variables. Interpret parameter estimates from the model.
b. Create an interaction model from the model in part (a). Let smoke be the moderator.
c. Create and interpret a plot to "probe" the interaction.
d. Create a regression of your own using the variables above (include one categorical variable). What did you find?

Assignment #8
STOR 455
Due July 24, 2017

Show only relevant output (i.e. relevant tables and plots). Print a paper copy to turn in during class. You may want to attach a copy of your SAS code. Do not answer questions in SAS code.

From the book

1. 10.5 a
2. 10.5 b
3. Obtain leverage values. Determine cutoff and create an index plot. What does it tell you?
4. 10.9 a
5. 10.9 e
6. 10.9 g

Assignment #9
STOR 455
Due July 26, 2017

Show only relevant output (i.e. relevant tables and plots). Print a paper copy to turn in during class. You may want to attach a copy of your SAS code. Do not answer questions in SAS code.

From the book

1. Refer to the dataset in problem 9.15
    a. Use the all-possible-regressions technique to determine the best model (consider multiple criteria – such as adjusted R^2, aic, …)
    b. Use stepwise regression with alpha limits of .1 and .15 to add or delete a variable, respectively
    c. Do results from part (a) and (b) compare?
2. Refer to the dataset in problem 11.6
    a. 11.6 a
    b. 11.6 b
    c. 11.6 c
    d. Use weighted least squares to perform the regression
    e. Compare estimates from part (a) and (d)

Ty Darnell
Stor 455
Assignment 3

2.1)
Looking at the CI for the slope, since the confidence interval does not include $B_1 = 0$ you can conclude that there is a linear association between Y and X. The implied level of significance is alpha = .05 since it is a 95% CI.

b) The scope of the model does not include 0 since it would be impractical for any district to have a population of 0. Thus the intercept does not have any intrinsic meaning of its own.

2.5)
a)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	-0.58016	2.80394	-0.21	0.8371	-5.29378	4.13347
copiers	1	15.03525	0.48309	31.12	<.0001	14.22314	15.84735

With 90% confidence we estimate that the mean service time increases by somewhere between 14.223 and 15.847 minutes for each additional copier serviced.

b)
Two-Sided Test
$H_0 = B_1 = 0$
$H_1 = B_1 \neq 0$
alpha = .1
t* = 15.03525/.48309 = 31.123
if p-value ≤ .1 reject H_0
if p-value > .1 fail to reject H_0
p-value < .0001
since p-value < alpha reject H_0 and conclude H_1 at the .1 level
conclude that there is a linear association between minutes and copiers serviced

c) The results from part a and b are consistent. The p-value is less than the significance level (0.1), which indicates that the results are statistically significant. The 90% confidence interval [14.223, 15.847] does not include the null hypothesis of $B_1 = 0$ which also means that the results are statistically significant.

2.14)

a)

Obs	copiers	minutes	yhat	stdp	lower	upper
1	6	.	89.631	1.39641	87.284	91.979

With 90% confidence we estimate that the mean service time is somewhere between 87.284 and 91.979 minutes when 6 copiers are serviced.

b)

Obs	copiers	minutes	pred	lower	upper	stdi
1	6	.	89.631	74.464	104.798	9.02223

With 90% confidence we predict that the service time on the next call in which 6 copiers are serviced will be somewhere between 74.464 and 104.798 minutes.

This is wider than the corresponding confidence interval. This is expected because the standard error for a prediction interval on an individual observation takes into account the uncertainty due to sampling and the variability of the individual observations around the predicted mean. The standard error for a confidence interval takes into account only the uncertainty due to sampling. Thus there is more variation in a prediction interval, making it wider.

2.24)
b)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	76960	76960	968.66	<.0001
Error	43	3416.37702	79.45063		
Corrected Total	44	80377			

$H_0 : B_1 = 0$
$H_1 : B_1 \neq 0$
alpha = .1
F* = MSR/MSE = 76960/79.45 = 968.66

If $p \leq$ alpha Reject H_0
If $p >$ alpha Fail to reject H_0
p-value < .0001

.0001 < .1 Reject H_0
Conclude H_1 at the .1 level
Thus there is a linear association between time spent and number of copiers serviced.