University of Guelph
STAT*3240 F19
Statistical Computing Lab 1 (and Assignment 1)

Hand in your submission to me by 1:00 pm on Friday September 20. Late assignments will not be accepted. You may slide your assignment under my door (550 MacN) if I am not there.

The room for the lab component of this course is SSC 1305 (and 1303, in the event of overflow). I will put the lab documents and data up on courselink, so you can do the labs and work on the assignments anywhere you have access to SAS and R.

Lab 1 will be fairly short; there will be a brief intro to SAS and we will use SAS to run a simple linear regression and create a couple of plots, use R to run a simple linear regression and create a plot, and get a start on Assignment 1. It would be best to get the computer portion of parts 1 and 2 out of the way during lab, then work on answering the questions later.

This lab will help you learn how to:

- Access SAS and run a some SAS code.
- Use SAS with the INFILE statement to access previously stored data.
- Use options in the SAS model statement to obtain residuals and predicted values.
- Import data into R and carry out a linear regression.

1 SAS

1.1 Accessing SAS, creating a data set, and running a regression program.

1. Open SAS version 9.4 by going to START > All Programs > SAS > SAS 9.4 (English).

2. Experiment with the SAS window features: Press the F7, F6, F5 function keys in succession to locate the Output, Log and Program editor windows. You can also find these windows using the View or Window drop down menus on the SAS menu bar.

   Output window: will contain output produced by the procedures you select.
   Log window: monitors the progress of the session. It will indicate any errors in your program.
   Editor window: This is where you will key in your SAS code.

3. Key the following program (or copy and paste) into the program editor window and submit it by clicking on the "running man" icon (located near the end of the top toolbar of SAS).

Note that all SAS code lines must end with a semicolon, unless it is a comment: /*Material within these symbols is ignored*/

One can also comment out a line by including an asterisk before the line e.g. *model y=x would be ignored by SAS (but you still need the semicolon).

Start copying here:

```
data lamename;
input x y;
datalines;
2 10
5 12
3 15
;
run;
proc reg; /* set-up to run a regression procedure */
model y=x;
run;
```

(Instead of `datalines;`, we could have used the equivalent command `cards;`, which is a throwback to computer punch cards. Yes, SAS goes way back to the 70s. James Goodnight, the CEO of SAS, is a billionaire and the richest person in the world that has a PhD in statistics!)

Run these commands by clicking `Run>Submit`

If all goes well, the regression output should appear in the results viewer.

To see what the data looks like, we can use the FSVIEW command in the little box in the top left corner:

```
FSVIEW work.lamename
```

It's a good idea to use FSVIEW to check to see that the data set is what you think it is, especially if you have imported it from a file.

(To clear all text in the log, output, or editor window, highlight the window and press ctrl e, or you can use `Edit > Clear all`.)

If the commands in the editor window disappear, then they can be retrieved by hitting the F4 key.

Let's get SAS to give us a plot of Y vs X for this data.

Key in the commands:

```
plot  y*x;
run;
```

Submit these statements (click the running man).

1.2 Using SAS to obtain residuals and predicted values

SAS includes various residual plots as part of the default output. We can also get output of residuals and predicted values by including the `r` and/or `p` options in the model statement:

```
model y=x/r p;
```

Depending on the output view, plots of the residuals are sometimes output as a default. We could also obtain a plot of the residuals using the plot statement:

```
plot r.*x;
run;
```

1.3 Using SAS with the INFILE statement to access previously stored data

Download the `3240_F19_A1_DDT.txt` data set from courselink and save it to your machine (remember where you saved it – you will need the path to the file in a minute).

This data set represents measurements of eggshell thickness (mm) and DDT levels (ppm) for 35 pelicans on Anacapa island in California. (We briefly discussed this data during the first lecture – see the notes for the reference. The 35 observations are a subset of the original data set.)

We could punch in the 35 pairs of observations, but that doesn't sound like too much fun and there is a good chance we'd make an error. A much better way to handle larger data sets is to import them from a file using the `infile` statement.

You will need to change the path to reflect where you have saved the file.

```
data DDT_eggs;
infile 'C:\STAT3420_F19\data\3240_F19_A1_DDT.txt';
input thickness DDT;
run;
```

N.B. SAS assigns names to the variables in the data set according to the order in which they appear in the input statement. (The names in the file do not matter.) Here, the variable in the first column will be given the name `thickness`, and the variable in the second column will be given the name `DDT`.

```
input thickness DDT;
```

is fundamentally different from

```
input DDT thickness;
```

Use FSVIEW to check to see that the data set has imported correctly: FSVIEW work.DDT_eggs

Now include the appropriate proc reg and model statements to run a regression with thickness as the response variable, and DDT as the explanatory variable. Also plot the residuals against DDT and include that plot with your assignment.

Q1 of Assignment #1:

(a) Attach the relevant SAS output. Include a scatterplot of thickness vs DDT, and a plot of residuals vs DDT.
(b) Interpret the parameter estimates of the slope and intercept in a clear and concise way.
(c) Carry out a t test of the null hypothesis that there is no linear relationship between eggshell thickness and DDT. Give appropriate hypotheses (in words and symbols), test statistic, p-value and conclusion.
(d) Calculate a 95% confidence interval for the true slope and give a proper interpretation of this interval.
(e) Carry out a t test of the null hypothesis that the true intercept is 0. Give appropriate hypotheses (in words and symbols), test statistic, p-value and conclusion. Is this a meaningful test?
(f) Do the residual plots give any indication that the assumptions of the model are violated? Justify your response.

2 R

Now let's take a quick look at how we'd carry out a regression analysis in R (another very commonly used statistical computing package). I am assuming you've been introduced to R in both 2040 and 2050, so this introduction will be brief.

Open R (Start > All Programs > R 3.6.1)

Fink et al (2007) investigated a possible relationship between the hand grip strength of young men and their facial attractiveness as perceived by young women. 14 male student volunteers had their hand grip strength measured in kilograms force (kgf), and their facial attractiveness was assessed by female students who viewed pictures of the faces of the men on a computer. The data file 3240_F19_A1_grip contains the handgrip and facial attractiveness measures for the 14 men in the study (we are using a subset of the data set from the original study).

Download the file 3240_F19_A1_grip.csv from courselink.

Now we'll import the file into R. There are many ways to do this, but a simple one is to use the read.csv command with the file.choose option, saving the result as an R object:

```
callthisRobjectwhateveryoufeellikebutthisnameisprobablytoolong<-read.csv(file.choose())
```

browse for the data file on your machine and double click it. It will be saved in whatever name you have chosen. From here on, I will assume you used the name `grip_data`.

Let's attach the data set (which will put it in R's search path, so the variables will be easily found):

```
attach(grip_data)
```

We are going to run a regression with facial attractiveness (called attractiveness in the data set) as the response variable, and grip strength (called grip in the data set) as the explanatory variable.

But first let's plot the data:

```
plot(grip,attractiveness)
```

The command:

```
grip_out<-lm(attractiveness~grip)
```

will run a linear regression (a Linear Model) of attractiveness on grip, and save the results to the object grip_out.

`summary(grip_out)` will give a summary of the results of the linear regression.

Now let's create a scatterplot, superimpose the least square regression line, and verify that the least squares regression line passes through the point $(\bar{X}, \bar{Y})$. Use the following 3 commands in order:

`plot(grip,attractiveness)` [scatterplot of attractiveness vs grip strength].

`abline(grip_out)` [superimposes the line from grip_out, which is the least squares line]

`points(mean(grip),mean(attractiveness),pch=16,cex=3)` [superimposes a point at $(\bar{X}, \bar{Y})$. The pch=16 makes it a solid circle, and the cex = 3 option makes it a little bigger than the default]

Include this plot in Assignment #1.

Part 2 of Assignment #1:

(a) Include the R output for the linear regression analysis.

(b) Include the scatterplot with superimposed regression line and point at ($\bar{X}$, $\bar{Y}$).

(c) Carry out a t test of the null hypothesis that there is no linear relationship between attractiveness and handgrip strength. Give appropriate hypotheses (in words and symbols), test statistic, p-value and conclusion.

3 Part 3: no computer required

(a) Show that $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{X}\sigma^2}{S_{XX}}$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least-squares estimators in our usual simple linear regression model.

Hint: There are different ways of going about it, but one way is to use this rule we discussed in class: If $Cov(Y_i, Y_j) = 0$ for $i \neq j$, then $Cov(\sum a_i Y_i, \sum b_i Y_i) = \sum a_i b_i Var(Y_i)$.

(b) For our usual simple linear regression model (under the usual assumptions), what is $Cov(Y_1, \bar{Y})$?

(c) Prove that the residuals sum to 0 (in least squares linear regression where the model includes an intercept). That is, prove that $\sum_{i=1}^{n} e_i = 0$.

(d) Suppose in a simple linear regression model it is known that $\beta_1 = 5$, and thus the model becomes $Y = \beta_0 + 5X + \epsilon$.

(i) Derive the least-squares estimator of β_0.

(ii) Assuming $Var(\epsilon) = \sigma^2$ for all X, derive the variance of the sampling distribution of $\hat{\beta}_0$ for this model.

(e) Suppose in a simple linear regression setting there are 3 possible X values: 1, 2, and 3. In reality, when $X = 1$, the Y variable has a normal distribution with a mean of 4 and a standard deviation of 1. When $X = 2$, the Y variable has a normal distribution with a mean of 8 and a standard deviation of 2. When $X = 3$, the Y variable has a normal distribution with a mean of 12 and a standard deviation of 4.

(a) What are the values of the parameters β_0 and β_1?

(b) What assumption of the linear regression model has been violated here?

4 Extra questions that won't be graded, but you should know how to do

- Show that when we use the least-squares estimation method in a simple linear regression model, the sample intercept ($\hat{\beta}_0$) is an unbiased estimator of the population intercept. That is, show that

$E(\hat{\beta}_0) = \beta_0$. (You may use the fact that the sample slope is an unbiased estimator of the population slope.)

- Suppose in a simple linear regression model it is known that $\beta_0 = 0$, and thus the model becomes $Y = \beta_1 X + \epsilon$ (with the usual assumptions on the ϵ term). Prove or disprove: for this model, the residuals always sum to 0.

- Show that $Cov(\bar{Y}, \hat{\beta}_1) = 0$.

- Prove that $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i - \bar{X})Y_i$.