Statistics 101A Lecture 1 Due Friday Jan. 17, 2020 at 5:00 PM
Homework One
Winter 2020

Data set: North Carolina Birth Data (NCBirthNew)
First of all, download the data from ccle week 1.
Data Size: `10000  132`
Variables Descriptions are posted on a separate file Week 1.
Note: (Use ggplot2 library for plots)

Problem One.

a) Create a histogram for the attribute "Birth Weight (g)" and test the claim that the average Birth Weight is 4300 g.
b) Recode the variable Gender of Child using Male instead of "1" and Female instead of "2"). "save it as GenderNew". Create a barplot for the GenderNew variable and test the claim that the proportion of Males is 0.50.
c) Construct a 95% confidence interval for the average Birth Weight (g)
d) Construct a 90% confidence interval for the proportion of Male babies in the data.

Problem Two:

a) Create a side-by-side box plot of the variable Birth Weight (g) of the two types of MomTran.
b) Conduct a two-tailed t-test comparing the average Birth Weight (g) of a Transferred Mom vs the average Birth Weight (g) of a Non-Transferred Mom. Report your p-value. (Assume Equal Variances).
c) Conduct a simple linear regression using Birth Weight (g) as your response variable and Gest Age (BC) as your predictor.
d) Create a scatter plot Gest Age (BC) vs Birth Weight (g) then plot the least square regression line on the same graph.
e) Report the summary of your linear model, interpret the slope and the y-intercept in the model based on the context.
f) Construct a 95% confidence interval for both: the slope and the y-intercept.
g) Using R or a calculator of your choice to calculate SST (total), SSE (residual), $SS_{Regression}$

Problem Three:

Use the SLR in Problem two to:
a) Compute a 95% confidence interval about the mean response for Gest Age (BC) = 20
b) Compute a 95% predication interval for a new observation when Gest Age (BC) = 20
c) Compare the two intervals.

Problem Four:

a) Conduct simple linear regression using Birth Weight (g) as outcome variable and MomTran as a predictor.
b) Create a scatter plot for the MomTran vs Birth Weight (g) then plot the least square regression line on the same graph.

c) Report the summary of your linear model, interpret the slope and the y-intercept in the model.
d) Compare the summary of your SLR in part c with the results of the t-test in Question Two Part (b). State your concludes?

Problem Five:
Below are some statistical summaries of the two variables "Gest Age (BC)" as the predictor and "Birth Weight (g)" as the response.

```
> summary(`Gest Age (BC)`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  16.00   38.00   39.00   38.43   40.00   44.00       2
> summary(`Birth Weight (g)`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  113.5  2951.0  3319.9  3258.5  3660.4  5334.5       2
> sd(`Gest Age (BC)`)
[1] NA
> var(`Gest Age (BC)`,na.rm=T)
[1] 5.928025
> var(`Birth Weight (g)`,na.rm=T)
[1] 394356.1
```

The sample size is `10000-2 = 998 (the 2 missing values are not considered in the SLR calculations)`

a) Use the statistical summaries to calculate S_{xx}, S_{xy}, S_{yy}=SST
b) Calculate the covariance between "Gest Age (BC)" and "Birth Weight (g)"
c) Calculate the linear correlation coefficient between "Age" and "Birth Weight (g)"
d) What are the values of slope and the y-intercept values of the SLR using "Gest Age (BC)" as the predictor and "Birth Weight (g)" as the response?
e) Use the equation of the SLR to predict the "Birth Weight (g)" of an infant with 40 weeks Gest Age (BC).

Good Luck