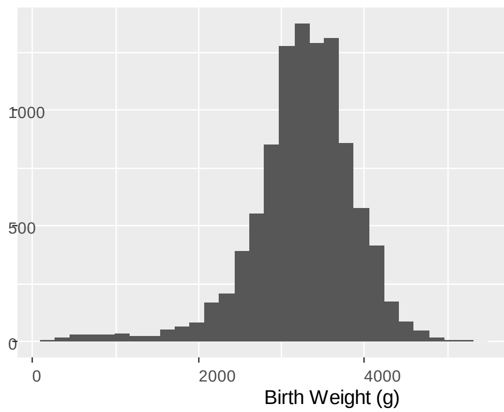
Statistics 101A Lecture 1 Due Friday Jan. 17, 2020 at 5:00 PM
Homework One
Winter 2020

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)

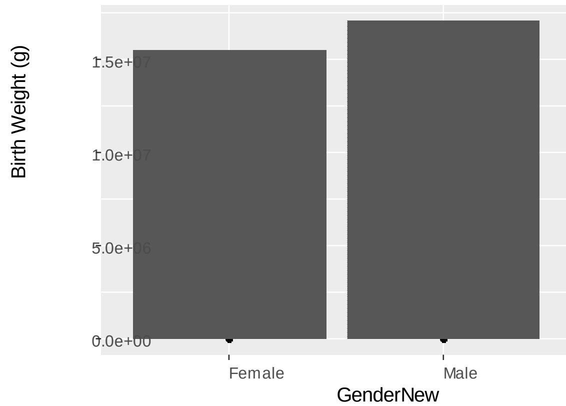
Problem One.

a) Create a histogram for the attribute "Birth Weight (g)" and test the claim that the average Birth Weight is 4300 g.




```
> 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(`Birth Weight (g)`,na.rm=T)
[1] 627.9778
> t.test(`Birth Weight (g)`,mu=4300,data=NCB)

	One Sample t-test

data:  Birth Weight (g)
t = -165.83, df = 9997, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 4300
95 percent confidence interval:
 3246.222 3270.844
sample estimates:
mean of x 
 3258.533 
```

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.




```
> table(GenderNew)
GenderNew
Female   Male
  4841   5159
> prop.table(table(GenderNew))
GenderNew
Female   Male
0.4841 0.5159
> prop.test(length(GenderNew[GenderNew=="Male"]),length(GenderNew),p=0.5)

        1-sample proportions test with continuity correction

data:  length(GenderNew[GenderNew == "Male"]) out of length(GenderNew), null
probability 0.5
X-squared = 10.049, df = 1, p-value = 0.001524
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5060509 0.5257368
sample estimates:
     p
0.5159

> binom.test(length(GenderNew[GenderNew=="Male"]),length(GenderNew),p=0.5)

        Exact binomial test

data:  length(GenderNew[GenderNew == "Male"]) and length(GenderNew)
number of successes = 5159, number of trials = 10000, p-value = 0.001523
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5060517 0.5257390
sample estimates:
probability of success
                0.5159
```

c) Construct a 95% confidence interval for the average Birth Weight (g)

```
95 percent confidence interval:
 3246.222 3270.844
sample estimates:
mean of x
 3258.533
```

d) Construct a 90% confidence interval for the proportion of Male babies in the data.

```
95 percent confidence interval:
 0.5060509 0.5257368
sample estimates:
```

```
     p
0.5159
```

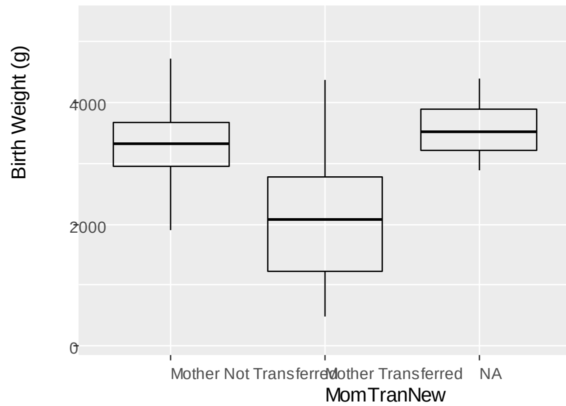
Or

```
95 percent confidence interval:
 0.5060517 0.5257390
sample estimates:
probability of success
                0.5159
```

Problem Two:

a) Create a side-by-side box plot of the variable Birth Weight (g) of the two types of MomTran.

```
> MomTranNew<-ifelse(MomTran==1,"Mother Transferred","Mother Not Transferr
ed")
> table(MomTranNew)
MomTranNew
Mother Not Transferred     Mother Transferred
                  9924                     71
```



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).

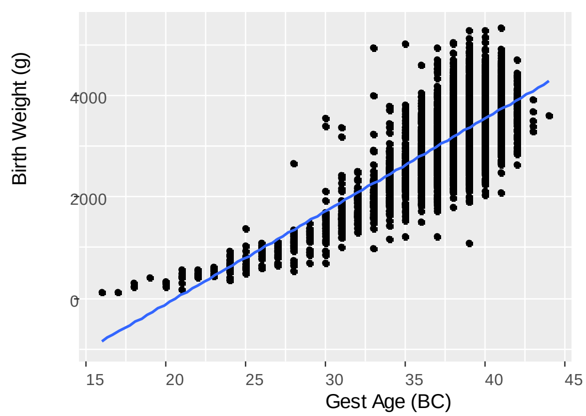
```
> t.test(`Birth Weight (g)`~MomTranNew,var.equal=T,data=NCB)

	Two Sample t-test

data:  Birth Weight (g) by MomTranNew
t = 16.153, df = 9992, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1047.966 1337.443
sample estimates:
mean in group Mother Not Transferred     mean in group Mother Transferred
                            3266.877                             2074.173
```

c) Conduct a simple linear regression using Birth Weight (g) as your response variable and Gest Age (BC) as your predictor.

```
> NCm1<- lm(`Birth Weight (g)` ~`Gest Age (BC)`)
> summary(NCm1)

Call:
lm(formula = `Birth Weight (g)` ~ `Gest Age (BC)`)

Residuals:
    Min      1Q  Median      3Q     Max
-2284.0  -297.7   -17.1   269.8  2672.3

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)      -3770.069     70.255  -53.66   <2e-16 ***
`Gest Age (BC)`   182.880      1.824  100.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 443 on 9995 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.5014,   Adjusted R-squared:  0.5013
F-statistic: 1.005e+04 on 1 and 9995 DF,  p-value: < 2.2e-16
```

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.

```
> NCm1$coefficients
    (Intercept) `Gest Age (BC)`
     -3770.0689        182.8796
```

f) Construct a 95% confidence interval for both: the slope and the y-intercept.

```
> confint(NCm1)
                    2.5 %     97.5 %
(Intercept)      -3907.7832 -3632.3545
`Gest Age (BC)`   179.3036   186.4555
```

g) Using R or a calculator of your choice to calculate SST (total), SSE (residual), $SS_{Regression}$

```
> # SST = Syy = (n-1)*var(Y)
> 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
> SST <- (9998-1)*var(`Birth Weight (g)`,na.rm = T)
> SST
[1] 3942377636
> SSE<- sum(NCm1$residuals^2)
> SSE
[1] 1961149716
> SSreg= SST - SSE
> SSreg
[1] 1981227920
```

Problem Three:

Use the SLR in Problem two to:
a) Compute a 95% confidence interval about the mean response for Gest Age (BC) = 20

```
new_data<-data.frame(Gest.Age..BC.=20)

predict(NCm1, data=NCB,newdata=new_data, interval = "confidence")
```

```
##      fit         lwr         upr
## 1    -112.4778   -178.9687   -45.98698
```

b) Compute a 95% predication interval for a new observation when Gest Age (BC) = 20

```
predict(NCm1, data=NCB,newdata=mdata, interval = "prediction")
```

```
##      fit         lwr         upr
## 1    -112.4778   -983.3097   758.354
```

c) Compare the two intervals.
We notice that those two intervals have the same fit (which is the midpoint), yet the prediction interval is much wider than the confidence interval.

Problem Four:

a) Conduct simple linear regression using Birth Weight (g) as outcome variable and MomTran as a predictor.

```
> NCm2<-lm(`Birth Weight (g)`~MomTranNew)
> summary(NCm2)

Call:
lm(formula = `Birth Weight (g)` ~ MomTranNew)

Residuals:
    Min      1Q  Median      3Q     Max 
-3153.4  -315.9    53.0   393.5  2295.6 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   3266.877      6.224  524.92   <2e-16 ***
MomTranNewMother Transferred -1192.704     73.839  -16.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 620 on 9992 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.02545,	Adjusted R-squared:  0.02535 
F-statistic: 260.9 on 1 and 9992 DF,  p-value: < 2.2e-16
```

b) Create a scatter plot for the MomTran vs Birth Weight (g) then plot the least square regression line on the same graph.

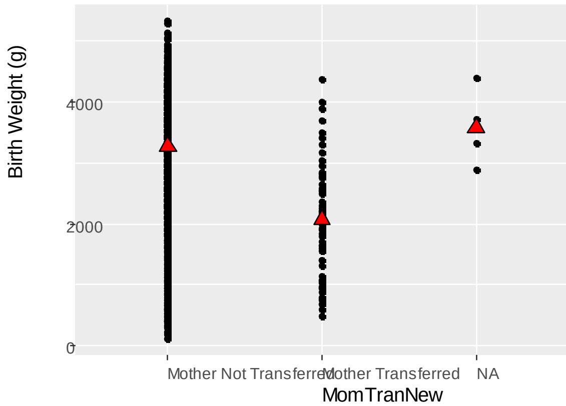
c) Report the summary of your linear model, interpret the slope and the y-intercept in the model.

```
> NCm2$coefficients
                 (Intercept) MomTranNewMother Transferred 
                    3266.877                    -1192.704 
```

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?

```
> t.test(`Birth Weight (g)`~MomTranNew,var.equal=T,data=NCB)

	Two Sample t-test

data:  Birth Weight (g) by MomTranNew
t = 16.153, df = 9992, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1047.966 1337.443
sample estimates:
mean in group Mother Not Transferred     mean in group Mother Transferred 
                            3266.877                             2074.173 
```

```
> qplot(MomTranNew,`Birth Weight (g)`,data=NCB)+ geom_smooth(method='lm')
```



The slope is 1192.70, which is the difference between the mean weight of two groups is 1192.70 grams. The y-intercept is `3266.877`, which us the mean of `group Mother Not Transferred`

We notice that the t-test and the SLR have the same p-value and t-value, and that the estimated slope is the same as the midpoint of the confidence interval in Question 2.

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 = 9998 (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

Sxx = 5.928025 × (9998 − 1) = 59262.465925

SST= Syy = 394356.1 × (9998 − 1) = 3942377931.7

Sxy = Sxx × β = 59262.465925 × 182.880 = 10837919.8

b) Calculate the covariance between “Gest Age (BC)” and “Birth Weight (g)”

```
> cov(`Birth Weight (g)`,`Gest Age (BC)`,use="complete.obs")
[1] 1078.662
```

Or

Sxy/(n-1)

c) Calculate the linear correlation coefficient between “Age” and “Birth Weight (g)”

```
> cor(`Birth Weight (g)`,`Gest Age (BC)`,use="complete.obs")
[1] 0.7080693
```

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?

```
> NCm1$coefficients
    (Intercept) `Gest Age (BC)` 
     -3770.0689        182.8796 
```

e) Use the equation of the SLR to predict the “Birth Weight (g)” of an infant with 40 weeks Gest Age (BC).

```
new_data=data.frame(Gest.Age..BC.=40)
predict(NCm1, newdata=new_data,interval="confidence")

##          fit        lwr        upr
## 1   3545.113   3534.781   3555.445
```

The predicted birth weight is 3545.113 grams, and a 95% confident interval is (3534.781, 3555.445)