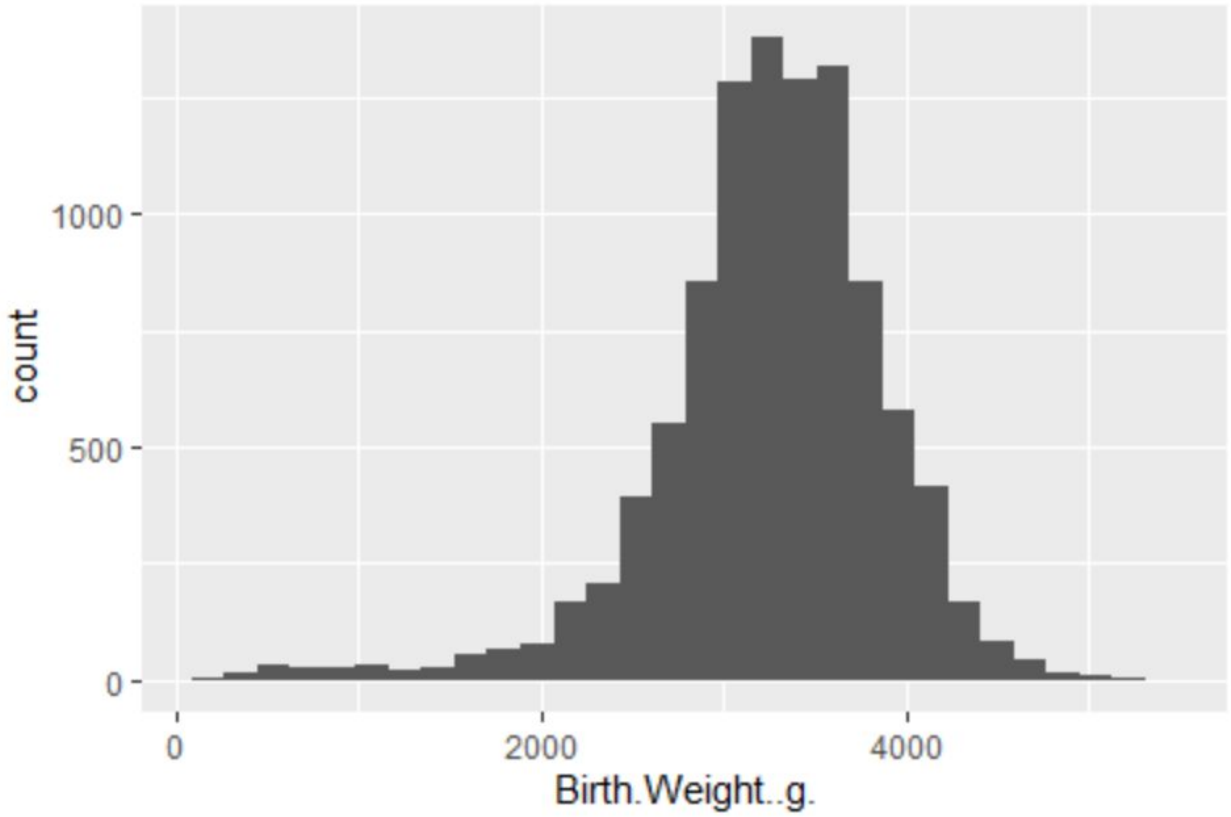
Leah Skelton
Statistics 101A
HW1
Due Friday Jan. 17, 2020 @ 5:00PM

Problem 1.

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

```
library(ggplot2)
data <- read.csv("C:/Users/choco/OneDrive/Documents/STATS101A/HW 1/NCBirthNew.csv"

#Problem 1
##A.
x <- data.frame(data$Birth.Weight..g.)
hist <- ggplot(data, aes(Birth.Weight..g.)) + geom_histogram()
summary(data$Birth.Weight..g.)
```



```
summary(data$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
```

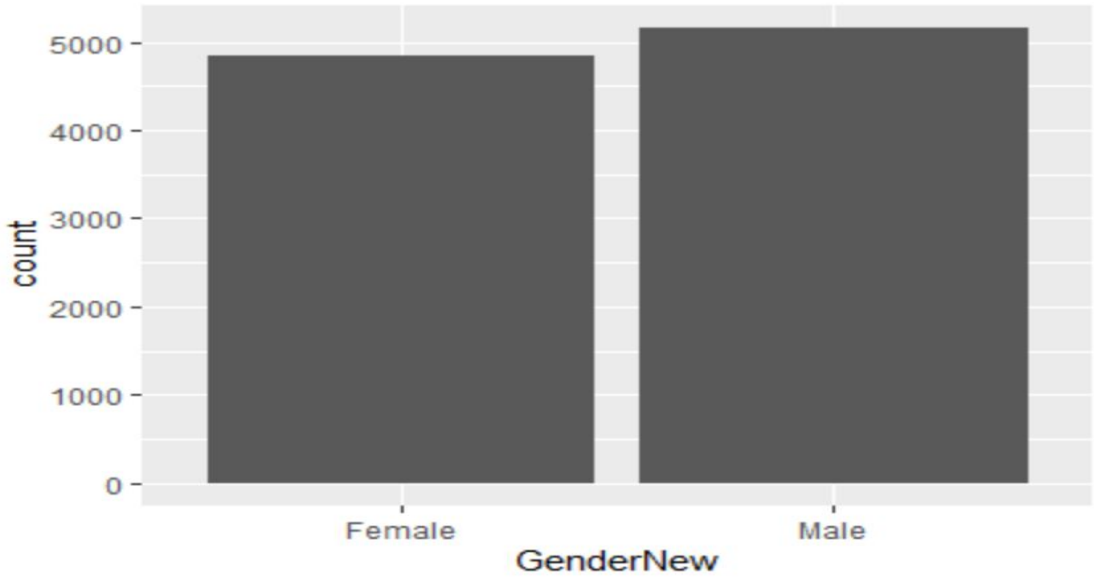
b) Recode the variable Gender of Child using Male instead of "1" and Female instead of "2"). "save it as GenderNew". Create a barplot for the GenderNew variable and test the claim that the proportion of Males is 0.50.

```
##B.
GenderNew <- vector(mode="character", length= length(data$Gender.of.child))
for (child in 1:length(data$Gender.of.child)){
  if (data$Gender.of.child[child] == 1){
    GenderNew[child] <- "Male"
  }
  else{
    GenderNew[child] <- "Female"
  }

}
```

```
GenderNew <- data.frame(GenderNew)

bar <- ggplot(GenderNew, aes(GenderNew)) + geom_bar()
```



```
n <- length(data$Gender.of.child)
sum_male <- sum(sum(data$Gender.of.child == 1))
proportion_male <- sum_male / n
```

```
> n <- length(data$Gender.of.child)
> sum_male <- sum(sum(data$Gender.of.child == 1))
> proportion_male <- sum_male / n
> proportion_male
[1] 0.5159
```

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

```
x <- qnorm(0.95)
sample_avg <- mean(data$Birth.Weight..g.,na.rm = TRUE)
standard_d <- sd(data$Birth.Weight..g.,na.rm = TRUE)
n <- length(data$Birth.Weight..g.)
margin_error <- x * standard_d / sqrt(n)

upper_b <- sample_avg + margin_error
lower_b <- sample_avg - margin_error
```

```
> upper_b
[1] 3268.862
> lower_b
[1] 3248.204
>
```

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

```r
##D.

library(DescTools)
males <- sum(sum(GenderNew == "Male"))
n <- length(data$Gender.of.child)
BinomCI(males, n , conf.level = 0.90, method = "clopper-pearson")
```

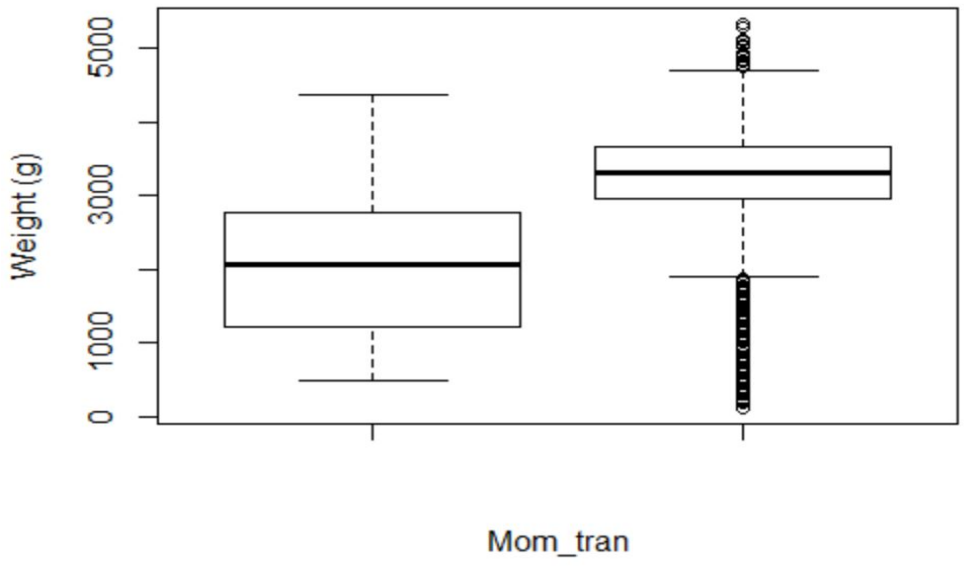
```
> BinomCI(males, n , conf.level = 0.90, method = "clopper-pearson")
        est    lwr.ci    upr.ci
[1,] 0.5159 0.5076272 0.5241659
```

Problem 2.

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

```r
#Problem 2
##A.
momtran_1 <- data$Birth.Weight..g.[data$MomTran == 1]
momtran_2 <- data$Birth.Weight..g.[data$MomTran == 2]

momtran <- c(momtran_1, momtran_2)
boxplot(momtran_1,momtran_2, horizontal = FALSE, ylab = "Weight (g)", xlab = "Mom_tran")
```



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).

```
##B.
test <- t.test(momtran_1,momtran_2, var.equal = TRUE)
```

```
        Two Sample t-test

data:  momtran_1 and momtran_2
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:
 -1337.443 -1047.966
sample estimates:
mean of x mean of y
 2074.173  3266.877
```

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

```
##C.
lin <- lm(data$Gest.Age ~ data$Birth.Weight..g., data=data)
summary(lin)
```

```
Call:
lm(formula = data$Gest.Age ~ data$Birth.Weight..g., data = data)

Coefficients:
          (Intercept)  data$Birth.Weight..g.
            30.094241                0.002589
```

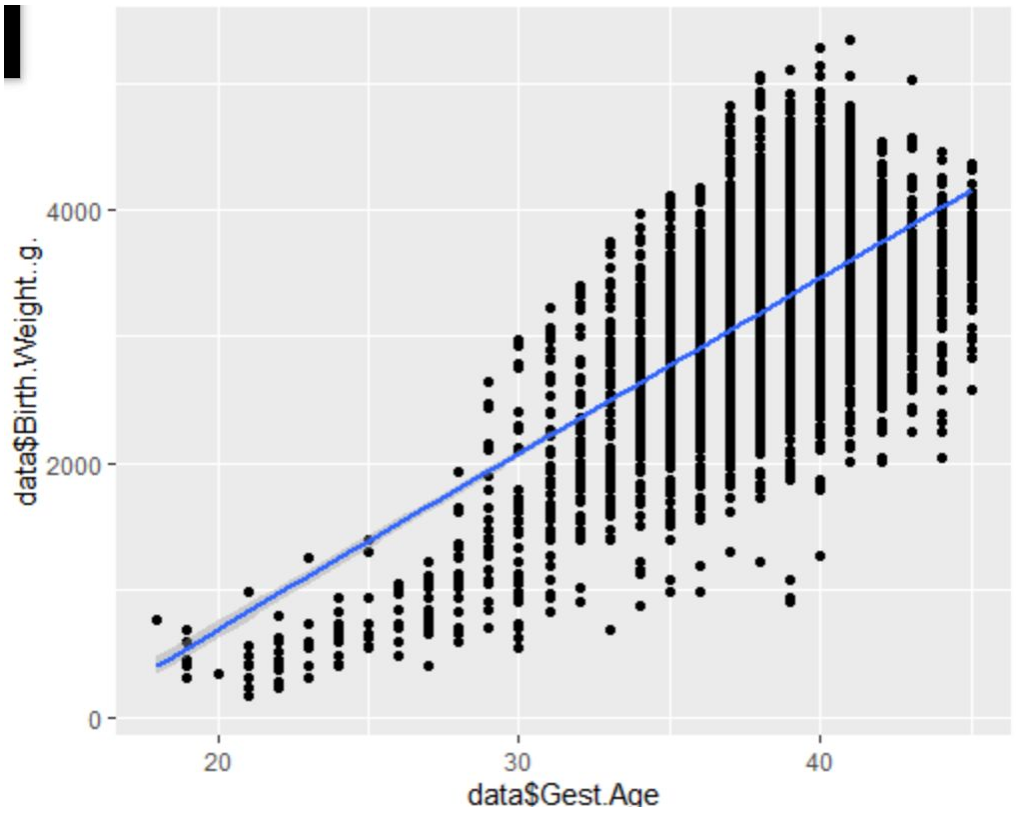
```
Residuals:
     Min       1Q   Median       3Q      Max
-14.0779  -1.1759   0.0445   1.2649   8.6160

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.009e+01  1.148e-01  262.04   <2e-16 ***
data$Birth.Weight..g. 2.589e-03  3.461e-05   74.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.163 on 9989 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.3591,    Adjusted R-squared:  0.3591
F-statistic:  5598 on 1 and 9989 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.

```
##D.
scatter <- ggplot(data,aes(data$Gest.Age, data$Birth.Weight..g.)) + geom_point()
+ geom_smooth(method='lm', formula= y~x)
```



e) Report the summary of your linear model, interpret the slope and the y-intercept in the model based on the context.

```
Call:
lm(formula = data$Gest.Age ~ data$Birth.Weight..g.)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0779  -1.1759   0.0445   1.2649   8.6160

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.009e+01  1.148e-01  262.04   <2e-16 ***
data$Birth.Weight..g. 2.589e-03  3.461e-05   74.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.163 on 9989 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.3591,    Adjusted R-squared:  0.3591
F-statistic:  5598 on 1 and 9989 DF,  p-value: < 2.2e-16
```

The intercept for this model is the estimate of the intercept, which would provide the expected birth weight in grams of a baby given that the Gest.Age is equal to zero. The slope for this model would be the square root of R-squared which is 0.599. This would mean that for every one-unit increase in Gest.Age, there's an expected average increase of 0.599 grams added to the baby's birth weight.

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

Slope:

```
##F.
upper_b <- (2.589*0.001) + 2*(3.461*0.00001)
lower_b <- (2.589*0.001) - 2*(3.461*0.00001)
```

```
> upper_b
[1] 0.00265822
> lower_b
[1] 0.00251978
```

Y-intercept:

```
##G.
upper_b <- 30.09 + 2*0.1148
lower_b <- 30.09 - 2*0.1148
```

```
> upper_b
[1] 30.3196
> lower_b
[1] 29.8604
```

g) Using R or a calculator of your choice to calculate SST (total), SSE (residual), SSRegression

```
> anova( lm(data$Gest.Age ~ data$Birth.Weight..g.))
Analysis of Variance Table

Response: data$Gest.Age
                       Df Sum Sq Mean Sq F value    Pr(>F)
data$Birth.Weight..g.    1  26198 26197.6  5597.7 < 2.2e-16 ***
Residuals             9989  46750     4.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SST = SSE + SSR

SSE is 26198, SSR is 46750, so SST is the addition of the two.

Problem 3.

a) Compute a 95% confidence interval about the mean response for Gest Age (BC) = 20

```
##A.
x <- qnorm(0.95)
sample_avg <- 20
standard_d <- sd(data$Gest.Age,na.rm = TRUE)
n <- length(data$Gest.Age)
margin_error <- x * standard_d / sqrt(n)

upper_b <- sample_avg + margin_error
lower_b <- sample_avg - margin_error
```

```
> upper_b
[1] 20.04451
> lower_b
[1] 19.95549
```

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

c) Compare the two intervals.

Problem 4.

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

```
##A.
lm <- lm(data$MomTran ~ data$Birth.Weight..g.)
```

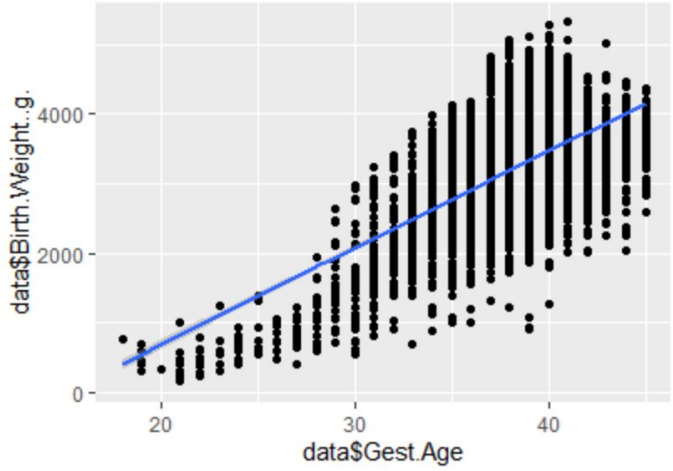
```
> lm <- lm(data$MomTran ~ data$Birth.Weight..g.)
> lm

Call:
lm(formula = data$MomTran ~ data$Birth.Weight..g.)

Coefficients:
          (Intercept)  data$Birth.Weight..g.
            1.923e+00              2.134e-05
```

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

```
##B.
scatter <- ggplot(data,aes(data$MomTran, data$Birth.Weight..g.)) + geom_point()
+ geom_smooth(method='lm', formula= y~x)
```



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

```
lm(formula = data$MomTran ~ data$Birth.Weight..g.)

Residuals:
     Min       1Q   Median       3Q      Max
-1.01661 -0.00147  0.00579  0.01306  0.07420

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.923e+00  4.383e-03  438.81   <2e-16 ***
data$Birth.Weight..g. 2.134e-05  1.321e-06   16.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08292 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
```

The intercept for this model is the estimate of the intercept, which would provide the expected birth weight in grams of a baby given that the Gest.Age is equal to zero. The slope for this model would be the square root of R-squared which is 0.1595. This would mean that for every one-unit increase in Gest.Age, there’s an expected average increase of 0.1595 grams added to the baby’s birth weight.

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?

The difference between the results here and the results in #2b is the difference means versus a correlation study. The problem in #2b would yield insight on the differences between two means, where this problem yields insight on how closely related these values are.

Problem 5.

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 Sxx, Sxy, Syy=SST

```
##5A.
x <- data$Gest.Age
y <- data$Birth.Weight..g.
n <- length(data$Gest.Age)

sxx <- sum(x^2) - sum(x)^2 / n
syy <- sum(y^2) - sum(y)^2 / n
sxy <- sum(x * y) - (sum(x) * sum(y)) / n
```

```
> sxx
[1] 85026
> syy
[1] 3963609464
> sxy
[1] 10491649
```

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

```
##B.
cov(data$Gest.Age, data$Birth.Weight..g., use = "complete.obs")
```

```
> cov(data$Gest.Age, data$Birth.Weight..g., use = "complete.obs")
[1] 1012.804
```

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

```
##C.
cor(data$Gest.Age, data$Birth.Weight..g., use = "complete.obs")
```

```
> cor(data$Gest.Age, data$Birth.Weight..g., use = "complete.obs")
[1] 0.5992756
```

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?

```
##D.
lm <- lm(data$Gest.Age ~ data$Birth.Weight..g.)
summary(lm)
```

```
Call:
lm(formula = data$Gest.Age ~ data$Birth.Weight..g.)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0779  -1.1759   0.0445   1.2649   8.6160

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.009e+01  1.148e-01  262.04   <2e-16 ***
data$Birth.Weight..g. 2.589e-03  3.461e-05   74.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.163 on 9989 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.3591,    Adjusted R-squared:  0.3591
F-statistic:  5598 on 1 and 9989 DF,  p-value: < 2.2e-16
```

The slope for this model would be the R value which is 0.5992. The y-intercept value would be 30.09.

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

```
##E.
B0 <- 30.09
x <- 0.5992
val <- 40

ans <- B0 + x * val
```

```
> B0 <- 30.09
> x <- 0.5992
> val <- 40
>
> ans <- B0 + x * val
> ans
[1] 54.058
```