HW 4 Due Monday Oct 2, 2017.
Upload R file to Moodle with name: HW4_490ID_YourClassID.R

Your classID: 52

Notice we are using the new system with your unique class ID. You should have received an email with
your unique class ID. Please make sure that ID is the only information on your hw that identifies you.

Do not remove any of the comments. These are marked by

################## Part 1: Linear Regression Concepts ######################

These questions do not require coding but will explore some important concepts.

"Regression" refers to the simple linear regression equation:
y = b0 + b1*x
This homework will not discuss other models.

1. (1 pt)
What is the interpretation of the coefficient B1?
(What meaning does it represent?)

Your answer here
Ans:
It is the value representing slope or the average increase in y associated with a one-unit
increase in x.

2. (1 pt)
Outliers are problems for many statistical methods, but are particularly problematic
for linear regression. Why is that? It may help to define what outlier means in this case.
(Hint: Think of how residuals are calculated)

Your answer here
Ans:
Outlier in this case may be a data with a a certain value of x placed inside the linear
regression with known b0 and b1 will result in a value y that would be a lot larger or
smaller than our original anticipated y value with the model. Thus, outliers in a data
may affect the model as the RSS takes in all values while computing, and any value that
differs alot from other values may change and affect the RSS value to represent the whole
data values incorrectly, causing deviation in results.

3. (1 pt)
How could you deal with outliers in order to improve the accuracy of your model?

Your answer here
Ans:
We can treat them depending on our model or different senarios. For example, we could
delete empty or missing values, assign the mean value of the data set to data that are
too large or too small, or treat them as values seperately.

################## Part 2: Sampling and Point Estimation ####################

The following problems will use the cats dataset and explore
the average body weight of female cats.

Load the data by running the following code

```
install.packages("MASS")
library(MASS)
data(cats)
```

4. (2 pts)
Subset the data frame to ONLY include female cats.

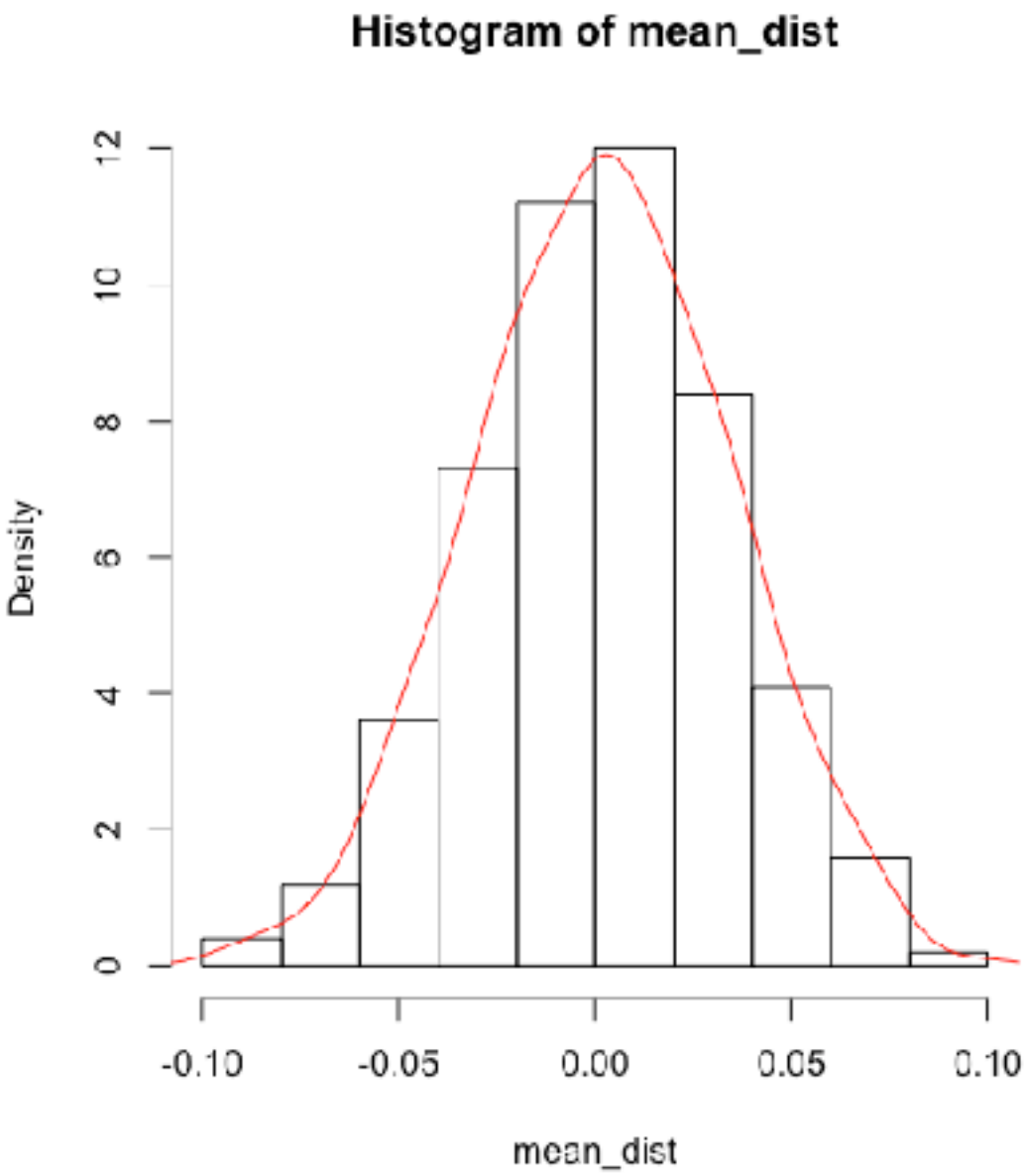
```
## Your answer here
cats[cats$Sex == "F",]
# Output:
# > cats[cats$Sex == "F",]
#    Sex Bwt  Hwt
# 1    F 2.0  7.0
# 2    F 2.0  7.4
# 3    F 2.0  9.5
# 4    F 2.1  7.2
# 5    F 2.1  7.3
# 6    F 2.1  7.6
# 7    F 2.1  8.1
# 8    F 2.1  8.2
# 9    F 2.1  8.3
# 10   F 2.1  8.5
# 11   F 2.1  8.7
# 12   F 2.1  9.8
# 13   F 2.2  7.1
# 14   F 2.2  8.7
# 15   F 2.2  9.1
# 16   F 2.2  9.7
# 17   F 2.2 10.9
# 18   F 2.2 11.0
# 19   F 2.3  7.3
# 20   F 2.3  7.9
# 21   F 2.3  8.4
# 22   F 2.3  9.0
# 23   F 2.3  9.0
# 24   F 2.3  9.5
# 25   F 2.3  9.6
# 26   F 2.3  9.7
# 27   F 2.3 10.1
# 28   F 2.3 10.1
# 29   F 2.3 10.6
# 30   F 2.3 11.2
# 31   F 2.4  6.3
# 32   F 2.4  8.7
# 33   F 2.4  8.8
# 34   F 2.4 10.2
# 35   F 2.5  9.0
# 36   F 2.5 10.9
# 37   F 2.6  8.7
# 38   F 2.6 10.1
# 39   F 2.6 10.1
# 40   F 2.7  8.5
# 41   F 2.7 10.2
```

```
# 42   F 2.7 10.8
# 43   F 2.9  9.9
# 44   F 2.9 10.1
# 45   F 2.9 10.1
# 46   F 3.0 10.6
# 47   F 3.0 13.0


## Use the sample function to generate a vector of 1s and 2s that is the same
## length as the subsetted data frame you just created. Use this vector to split
## the 'Bwt' variable into two vectors, Bwt1 and Bwt2.

## IMPORTANT: Make sure to run the following seed function before you run your sample
## function. Run them back to back each time you want to run the sample function to ensure
## the same seed is used every time.

## Check: If you did this properly, you will have 24 elements in Bwt1 and 23 elements
## in Bwt2.

set.seed(676)  ###########
## Your answer here
female_cat = cats[cats$Sex == "F",]
female_cat["sample12"] = sample(c(1,2), length(cats[cats$Sex=="F", 1]), replace = T)
Btw1 = female_cat$Bwt[female_cat$sample12 == 1]
Btw2 = female_cat$Bwt[female_cat$sample12 == 2]

Btw1
# > Btw1
# [1] 2.0 2.0 2.1 2.1 2.1 2.2 2.2 2.2 2.3 2.3 2.3 2.4 2.4 2.4 2.4 2.5 2.5 2.6 2.6 2.7 2.7 2.7
# [23] 2.9 3.0
Btw2
# > Btw2
# [1] 2.0 2.1 2.1 2.1 2.1 2.1 2.1 2.2 2.2 2.2 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.6 2.9 2.9
# [23] 3.0

## 5. (3 pts)  ##########
## Calculate the mean and the standard deviation for each of the two
## vectors, Bwt1 and Bwt2. Use this information to create a 95%
## confidence interval for your sample means (you can use the following formula
## for a confidence interval: mean +/- 2 * standard deviation).
## Compare the confidence intervals -- do they seem to agree or disagree?

## Your answer here
mean(Btw1)
# > mean(Btw1)
# [1] 2.4
sd(Btw1)
# > sd(Btw1)
# [1] 0.2734641
mean(Btw2)
# > mean(Btw2)
# [1] 2.317391
sd(Btw2)
# > sd(Btw2)
# [1] 0.2741137

## Ans:
# Confidence interval for Btw1: 2.4 - 2*0.2734641 ~ 2.4 + 2*0.2734641 = [1.85,2.947]
# Confidence interval for Btw2: 2.317391 - 2*0.2741137 ~ 2.317391 - 2*0.2741137 = [1.77,2.86]
# Both confidence intervals are similar as both values are very close. Thus they seem to agree.
```

6. (4 pts)
Draw 1000 observations from a standard normal distribution. Calculate the sample mean.
Repeat this 500 times, storing each sample mean in a vector called mean_dist.
Plot a histogram of mean_dist to display the distribution of your sample mean.
How closely does your histogram resemble this normal distribution? Explain.

Your answer here

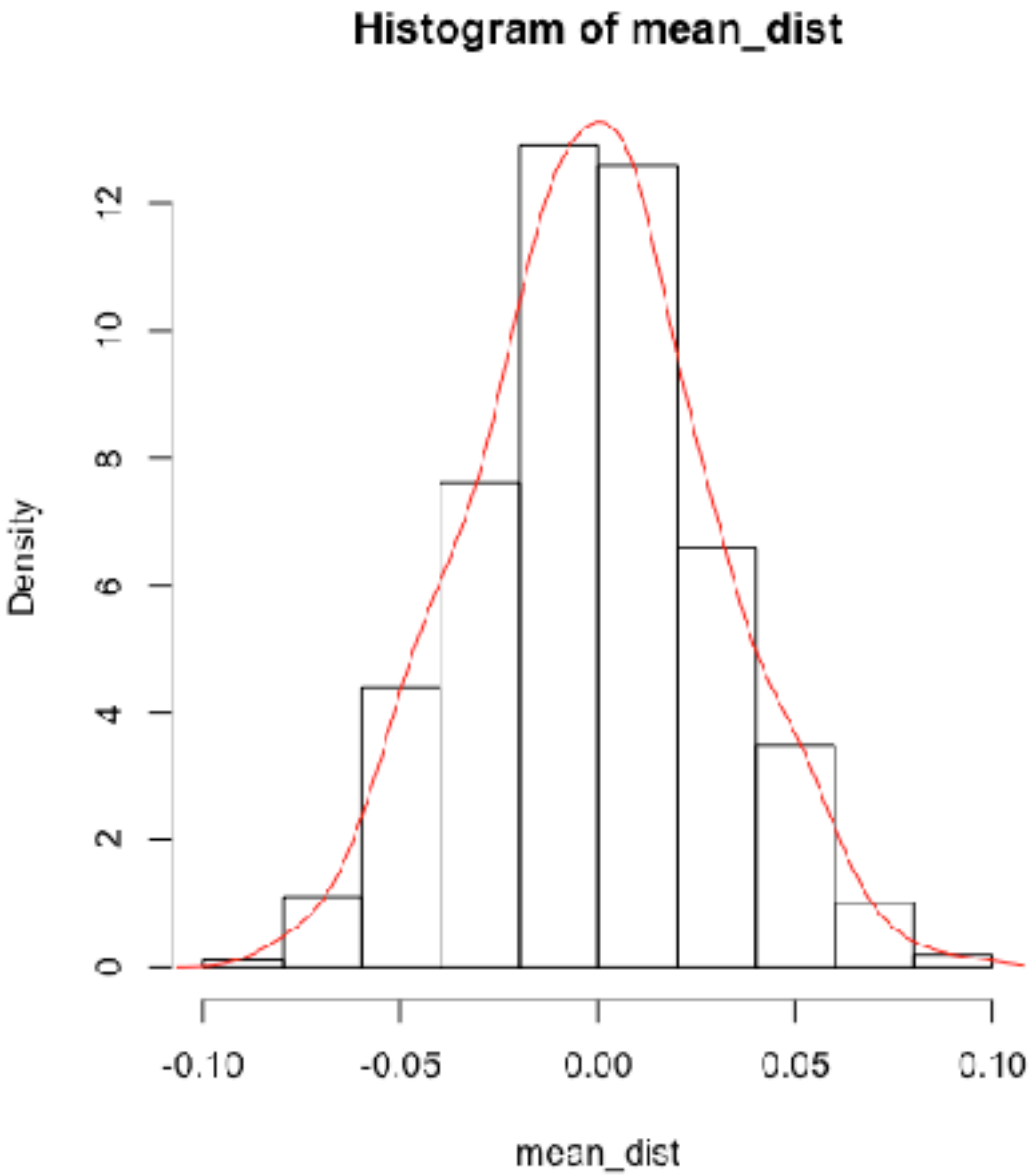
```
mean_dist = c()
for(i in 1:500){
  mean_dist = c(mean_dist, mean(rnorm(1000)))
}

hist(mean_dist, freq = F)
lines(density(mean_dist), col = "red")
```



Ans:
As we can see from the density line, which is bell shaped, resembles very closely to a normal distribution.

7. (3 pts)
Write a function that implements Q5.

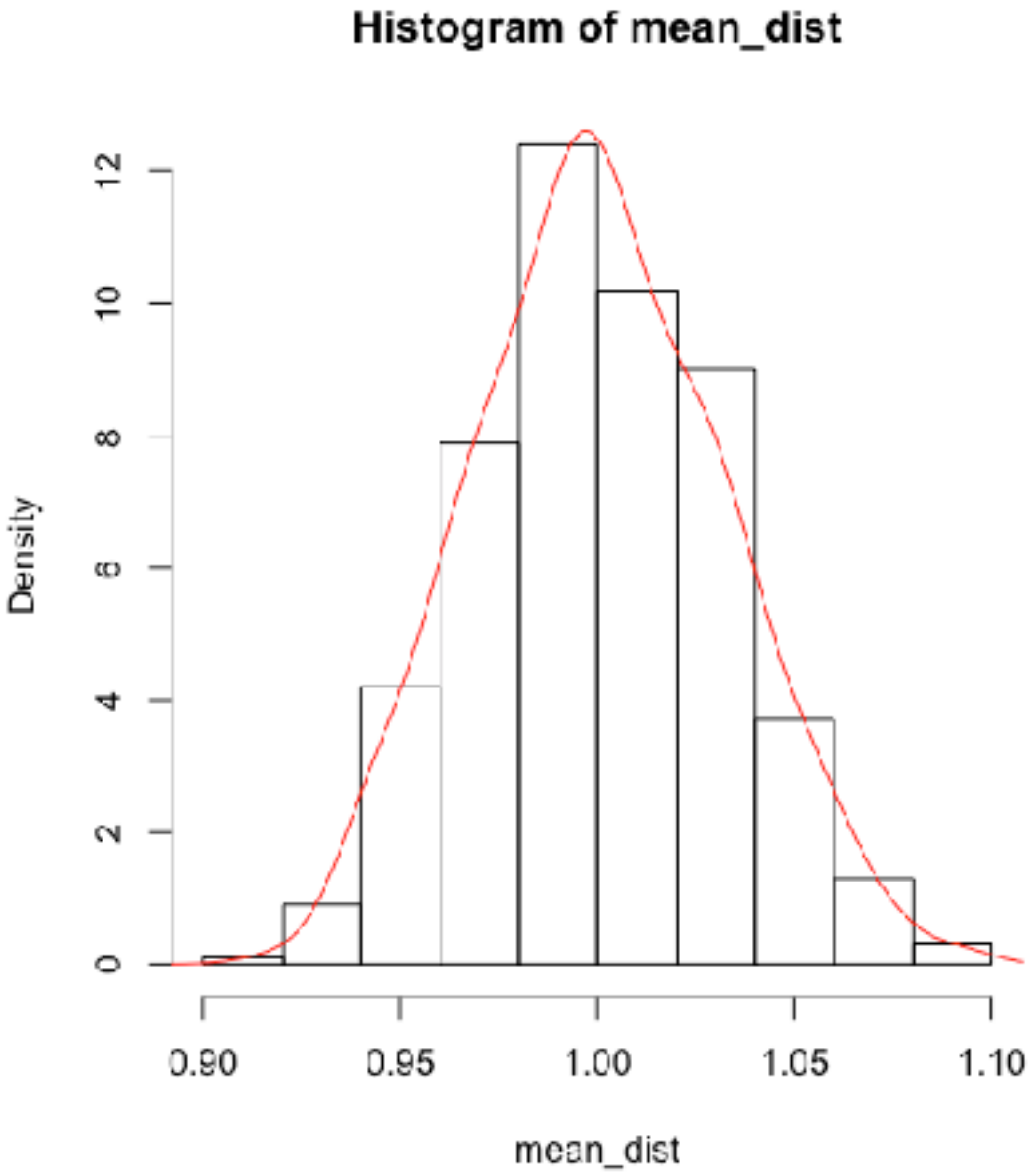
```
HW.Bootstrap=function(distn,n,reps){
  set.seed(666)

  ### Your answer here
  mean_dist = c()
  for(i in 1:reps){
    mean_dist = c(mean_dist, mean(distn(n)))
  }
  hist(mean_dist, freq= F)
  lines(density(mean_dist), col="red")
}

HW.Bootstrap(rnorm, 1000, 500)
```



Use the function you write to repeat the experiment in Q5 but instead of the
normal distribution as we used above, use an exponential distribution with mean 1.
Check your histogram and write out your findings.
(Hint: HW.Bootstrap(rexp,n,reps))

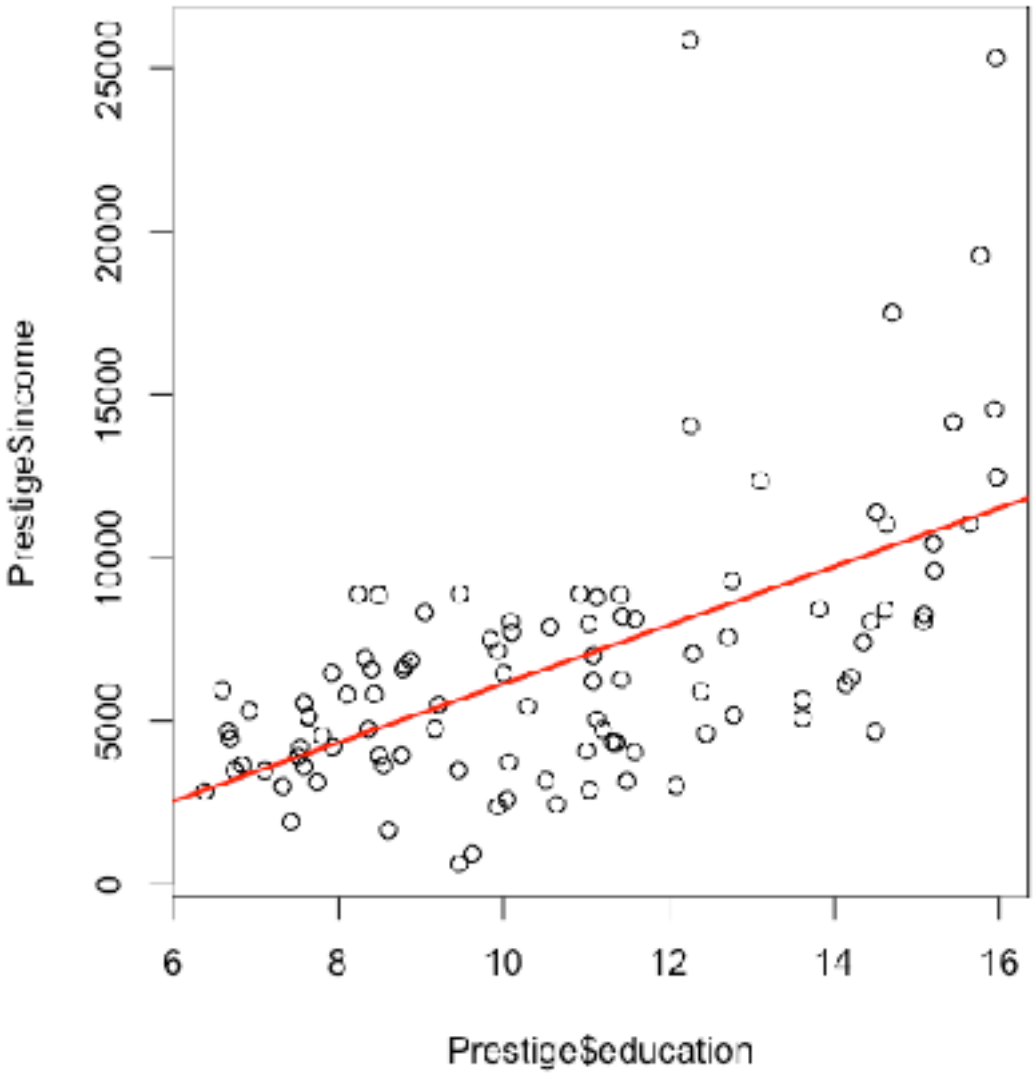
Your answer here

HW.Bootstrap(rexp, 1000, 500)



Ans:
My findings are that the distribution is also bell shaped, resembling a normal distribution.
Also, both histograms are somewhat similar.
However the difference is that the exponential distribution does not have any negative values.

```
################### Part 3: More Linear Regression ######################

## This problem will use the Prestige dataset.
## Load the data by running code below
install.packages("car")
library(car)
data(Prestige)

## We will focus on this two variables:
## income: Average income of incumbents, dollars, in 1971.
## education: Average education of occupational incumbents, years, in 1971

## Before starting this problem, we will declare a null hypthosesis that
## education has no effect on income .
## That is: H0: B1 = 0
##          HA: B1 != 0
## We will attempt to reject this hypothesis by using a linear regression


## 8. (2 pt)
## Fit a linear regression using of Prestige data using education to predict
## income, using lm(). Examine the model diagnostics using plot(). Would you
## consider this a good model or not? Explain.

## Your answer here
lm(Prestige$income ~ Prestige$education, data = Prestige)
# > lm(Prestige$income ~ Prestige$education, data = Prestige)
#
# Call:
#   lm(formula = Prestige$income ~ Prestige$education, data = Prestige)
#
# Coefficients:
#   (Intercept)  Prestige$education
# -2853.6              898.8
plot(Prestige$education, Prestige$income)
abline(lm(Prestige$income ~ Prestige$education, data = Prestige), col = "Red", lw = 2)
```



Ans:
The plot show there is positive correlation between income and education, I would
consider this a good method as the regression line also shows out clearly that
both values have a positive correlation that somewhat follows the line.

9. (2 pts)
Using the information from summary() on your model (the output from the lm() command), create a
95% confidence interval for the coefficient of education variable

Your answer here
summary(lm(Prestige$income ~ Prestige$education, data = Prestige))
# > summary(lm(Prestige$income ~ Prestige$education, data = Prestige))
#
Call:
lm(formula = Prestige$income ~ Prestige$education, data = Prestige)

```
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -5493.2 -2433.8   -41.9  1491.5 17713.1
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)         -2853.6     1407.0  -2.028   0.0452 *
#   Prestige$education    898.8      127.0   7.075 2.08e-10 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 3483 on 100 degrees of freedom
# Multiple R-squared:  0.3336,          Adjusted R-squared:  0.3269
# F-statistic: 50.06 on 1 and 100 DF,  p-value: 2.079e-10

# Ans:
# As a 95% confidence interval lies between two Standard Error length on both sides of the value
# the confidence interval for B0 and B1 are:
# B0 = -2853.6           +- 2*1407 ==> [-5667,-39.6],
# B1 = 898.8   +- 2*127 ==> [644.8,1152.8]



## 10. (2 pts)
## Based on the result from question 9, would you reject the null hypothesis or not?
## (Assume a significance level of 0.05). Explain.

## Your answer here

# Ans: Yes, as the p-value of the education (2.08e-10) is so small that it is close to zero,
# which is a lot more smaller than the typical cutoff of p-value which is 5%.
# Also the t-value is relatively far from zero. Both indicating that there is a relation
# between education and income, and that we can reject the null hypothesis.
# Thus, I would reject the null hypothesis.



## 11. (1 pt)
## Assuming that the null hypothesis is true.
## Based on your decision in the previous question, would you be committing a decision error?
## If so, which type of error?

## Your answer here

# Ans: Assuming that the null hypothesis is true. It would mean I have made a mistake of
# rejecting the null hypothesis when it is true. Thus I would be making a Type 1 Error,
# which is -- when a null hypothesis is true and we reject it, which would be exactly what
# I did for this particular case.



## 12. (1 pt)
## Discuss what your regression results mean in the context of the data.
## (Hint: Think back to Question 1)
```

```
## Your answer here
# Ans:
# Residuals:
#    Min      1Q  Median      3Q     Max
# -5493.2 -2433.8   -41.9  1491.5 17713.1
#
# Coefficients:
#    Estimate Std. Error t value Pr(>|t|)
# (Intercept)         -2853.6     1407.0  -2.028   0.0452 *
#    Prestige$education    898.8      127.0   7.075 2.08e-10 ***
#    ---
#    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 3483 on 100 degrees of freedom
# Multiple R-squared:  0.3336,          Adjusted R-squared:  0.3269
# F-statistic: 50.06 on 1 and 100 DF,  p-value: 2.079e-10

## Ans:
# As we can see from the regression results,
# from the education coefficient results, we can see that the slope is 898.8, meaning
# that income is estimated to increase by 898.8 dollars for every year of education on had.
# Also, this model lets us see that there is a correlation between education and income,
# the Residual standard error also gives us clues of our models quality,
# with a RSE of 3483, predicting the average amount of actual income will deviate
# from the true regression ine.
# And as for how well our model is fitting the actual data, from the two R-squared values,
# we can see that roughly around 32% of our response variable - income can be explained
# by our predictor variable - education.
```