Project1

Benjamin Osei Arthur

11/8/2022

Gender discrimination

Gender discrimination in bank salaries. In the 1970's, Harris Trust was sued for gender discrimination in the salaries it paid its employees. One approach to addressing this issue was to examine the starting salaries of all skilled, entry-level clerical workers between 1965 and 1975. The following variables, which can be found in `banksalary.csv`, were collected for each worker.

- `bsal` = beginning salary (annual salary at time of hire)
- `sal77` = annual salary in 1977
- `sex` = MALE or FEMALE
- `senior` = months since hired
- `age` = age in months
- `educ` = years of education
- `exper` = months of prior work experience

```
library(readr)
banksal <- read_csv("banksalary.csv")
# banksal <- as.tibble(banksal)
```

a. Identify observational units, the response variable, and explanatory variables. Observational units: Employees Response variable : beginning salary Explanatory variable : salary in 1977,sex,senior,age,education,experience

b. The mean starting salary of male workers ($5957) was 16% higher than the mean starting salary of female workers ($5139). Confirm these mean salaries. Compute the standard error and 95% confidence interval for this comparison. Discuss any possible sources of bias or unmodeled uncertainty. Is this enough evidence to conclude gender discrimination exists? Give reasons why or why not.

```
mu_male <- 5957
mu_female <- 5139
# Computing the mean and standard deviation
data_summary<-banksal %>%
  group_by(sex)%>%
  summarise(Means = mean(bsal),variance=var(bsal),std_deviation = sd(bsal),n=n())
  data_summary
```

```
## # A tibble: 2 x 5
##   sex    Means variance std_deviation     n
##   <chr>  <dbl>    <dbl>         <dbl> <int>
## 1 FEMALE 5139.  291460.          540.    61
## 2 MALE   5957.  477112.          691.    32
```

From the above computation the means for female and male are 5139 and 5957 respectively after rounding of which is equal to the means in the question.

The standard error for comparison of mean is

$$S.E = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

```
# let var_m is the male variance and var_f is the female variance.
m_mean =5957
f_mean =5139
var_f =291460.3
var_m = 477112.5
n_f =61
n_m =32
mean_diff = m_mean-f_mean
z_multiplier =qnorm(c(.025, .975), mean = 0, sd = 1)


SE = sqrt(var_f/n_f + var_m/n_m)
SE
```

```
## [1] 140.3132
```

```
CI = mean_diff + z_multiplier * SE
CI
```

```
## [1]  542.9911 1093.0089
```

The standard error is SE = 140.3132 and the 95% confidence interval the comparison is 542.9911 and 1093.0089. The mean difference value 818 which lies between the confidence interval. From the analysis , i can conclude that male have over 16% higher income than female. We have a small sample size of 93 to much a 10 years. So i think according to the data available , i would conclude that males have higher salaries , but i would suggest we get more data point to make this inference.

c. How would you expect age, experience, and education to be related to starting salary? Generate appropriate exploratory plots; are the relationships as you expected? Do you see any patterns of concern for modeling?

I would extrapolate and assume experience and education would have a strong positive relationship with the starting salary. To confirm my clam i will perform scatter plot of each variable with the beginning salary and also find the correlation among them.

```
sc_p1 <-banksal %>%
  ggplot(aes(bsal,age,color= sex)) + geom_point()+
  labs(title = "Relationship between ",
       subtitle = "Beginning salary and Age",
       x="Beginning salary",y="age")+theme(legend.position = "none")

sc_p2 <- banksal %>%
  ggplot(aes(bsal,exper,color= sex)) + geom_point()+
   labs(title = "Relationship between",
        subtitle = "Beginning salary and Experience",
       x="Beginning salary",y="Experience")

sc_p3 <-banksal %>%
  ggplot(aes(bsal,educ,color= sex)) + geom_point()+
   labs(title = "Relationship between ",
        subtitle = "Beginning salary and Education",
       x="Beginning salary",y="Education")
```

```
sc_p1 + sc_p2 / sc_p3
```

From the correlation below it shows a weak positive correlation between beginning salary and age, experience and education respectively.

```
cor(banksal[,c(1,4,5,6,7)])
```

```
##               bsal      senior         age        educ       exper
## bsal     1.00000000 -0.28584352  0.03389932  0.41198516  0.16674049
## senior -0.28584352  1.00000000 -0.18448263  0.05984385 -0.07466085
## age      0.03389932 -0.18448263  1.00000000 -0.22525298  0.79787476
## educ     0.41198516  0.05984385 -0.22525298  1.00000000 -0.10117309
## exper    0.16674049 -0.07466085  0.79787476 -0.10117309  1.00000000
```

d. Why might it be important to control for seniority (number of years with the bank) if we are only concerned with the salary when the worker started?

```
cor(banksal[,c(1,4)])
```

```
##               bsal     senior
## bsal     1.0000000 -0.2858435
## senior -0.2858435  1.0000000
```

From the table above,We must be concern with the seniority since it has a negative correlation with the beginning salary.

e. By referring to exploratory plots and summary statistics, are any explanatory variables (including sex) closely related to each other? What implications does this have for modeling?

```
cor(banksal[,c(1,2,4,5,6,7)])
```

```
##               bsal      sal77      senior         age        educ       exper
## bsal    1.00000000  0.4223695 -0.28584352  0.03389932  0.41198516  0.16674049
## sal77   0.42236948  1.0000000  0.12595515 -0.54674689  0.42102125 -0.37198640
## senior -0.28584352  0.1259551  1.00000000 -0.18448263  0.05984385 -0.07466085
## age     0.03389932 -0.5467469 -0.18448263  1.00000000 -0.22525298  0.79787476
## educ    0.41198516  0.4210213  0.05984385 -0.22525298  1.00000000 -0.10117309
## exper   0.16674049 -0.3719864 -0.07466085  0.79787476 -0.10117309  1.00000000
```

From the table above there exist a multicollinearity between age and experience, senior and experience which can cause a problem to the fitted model. Generally , the coefficient represents the mean change in dependent variable for a unit change in the independent variable, when other variable are constant. When an independent variable are strongly correlated , it mean the changes in one variable are associated with the shift in the other variable. This makes it more difficult for the model to estimate the relationship between each independent variable and the dependent variable independently because

(https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/)

f. Fit a simple linear regression model with starting salary as the response and experience as the sole explanatory variable (Model 1). Interpret the intercept and slope of this model; also interpret the R-squared value. Is there evidence of a relationship between experience and starting salary?

```
model <- lm(bsal ~ exper, data= banksal)
summary(model)
```

```
##
## Call:
## lm(formula = bsal ~ exper, data = banksal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1389.02  -503.33   -36.03   383.14  2740.08
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5289.0217   109.2984  48.391   <2e-16 ***
## exper          1.3009     0.8064   1.613     0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 703.5 on 91 degrees of freedom
## Multiple R-squared:  0.0278, Adjusted R-squared:  0.01712
## F-statistic: 2.602 on 1 and 91 DF,  p-value: 0.1102
```

The intercept 5289 is the expected beginning salary when experience is zero. The experience coefficient of regression equation is 1.3. This coefficient represents the mean increase of beginners salary for every additional months of experience prior to the work.If experience is increased by a month, the average beginners salary increases by 1.3 dollars. An r square of 0.0278 reveals that 2.8% of the data fit the regression model. and since r-square is a low value it indicate a bad fit for the model.The correlation between beginning salary and experience is 0.167 which suggest a weak or no relationship between them.

```
cor(banksal[,c(1,7)])
```

```
##            bsal     exper
## bsal  1.0000000 0.1667405
## exper 0.1667405 1.0000000
```

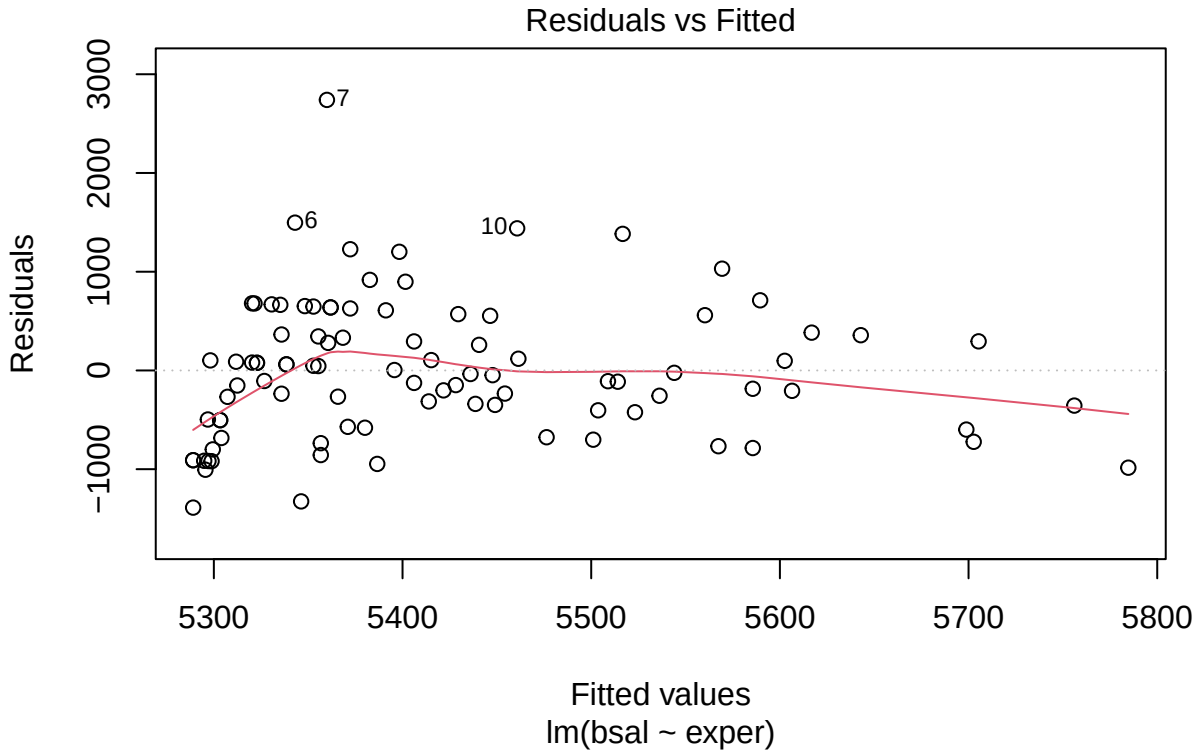
g. Does Model 1 meet all simple linear regression assumptions? List each assumption and how you decided if it was met or not.

Validity: Sampling beginning salary and experience to fit a regression model is not valid to determining the influence of gender discrimination against beginners salary. To make it valid we can add the sex predictor to the model and this would make a valid model answer our question.

Representatives: The sample data do not represent the interest of the population. Since the main research question is understand the gender discrimination of beginner salary at banks. From the sample data we have only two variable to make inference from which are beginner salary and experience. There is a weak positive relationship between them, but we can not conclude that an employees with an many months of experience would get a higher income because a new employee with education and less experience would get more income than an employee with more months of experience. Also we have to fact in the gender of the employees since its the research main interest. Therefore the sample data do not represent the population and research interest.

Additivity and linearity: The linearity of the model can be checked by inspecting the residual and fitted plot. From the plot, majority of the red line is off the zero point which indicate issue with linearity and also the points has a little linear trend and this suggests a problem with some aspect of the linear model, hence there is a nonlinear relationship between beginner salary and experience.

```
plot(model,1)
```



Independence of errors: To check for independence of error , we will apply the durbin watson test to check the null hypothesis ,which states that the errors are not auto-correlated with themselves. From the test, the pvalue 0.002 is less than 0.05, we reject the null hypothesis and concluded that the residuals are autocorrelated hence dependent.

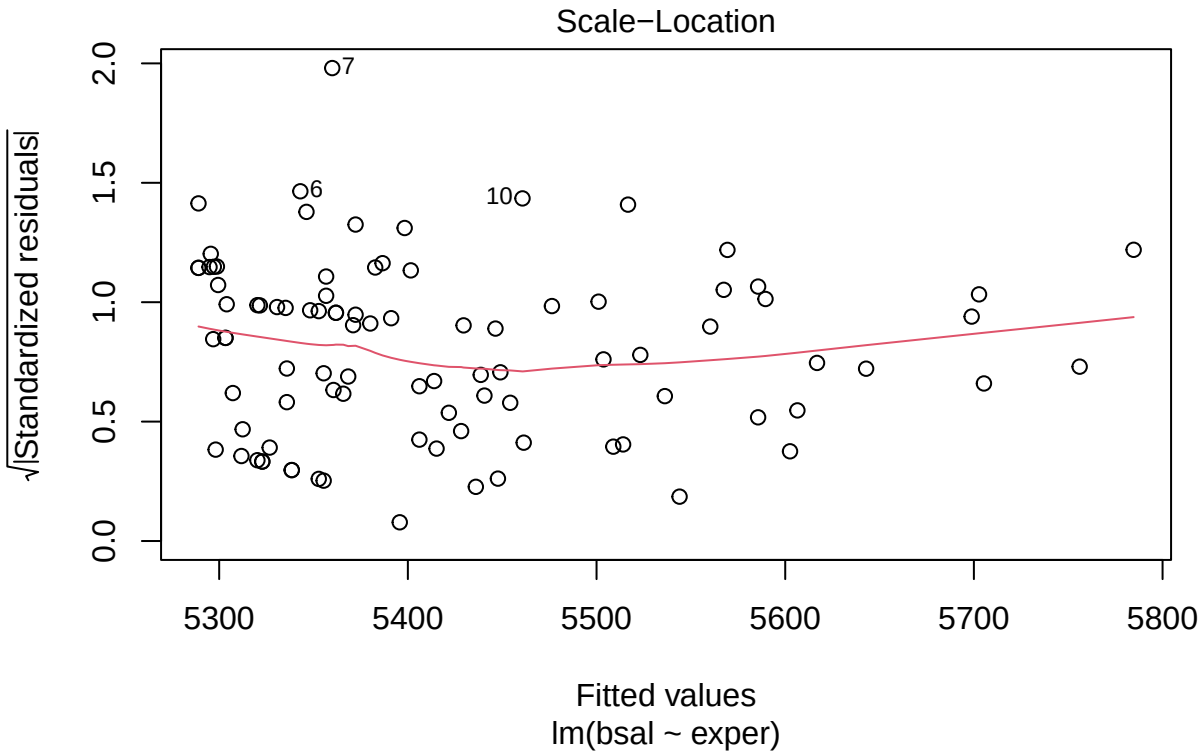
```
durbinWatsonTest(model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3095469      1.378102   0.002
##  Alternative hypothesis: rho != 0
```

Equal variance of errors (constant variance): we check the assumption of equal variance using the scale-location plot. From the plot, we can see a horizontal red line with equally spread points around it. its can further be

seen that there is no variability as residual points increase with the value of the fitted out come.

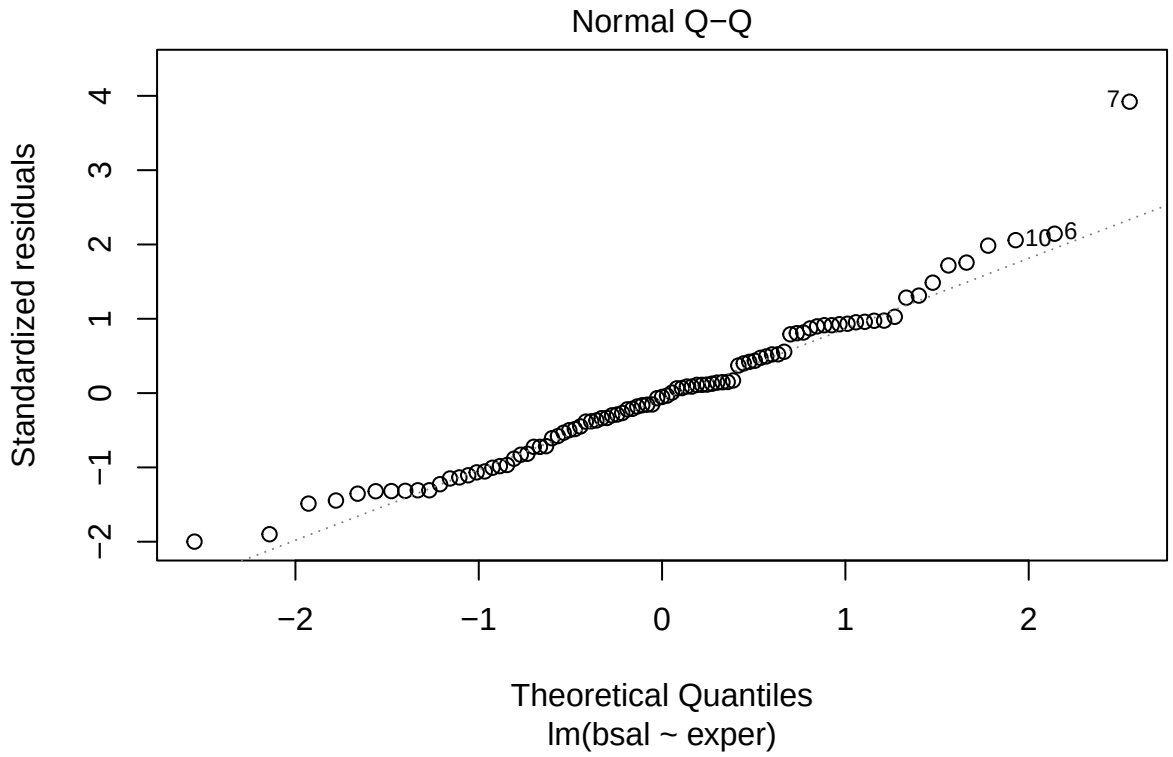
```
plot(model,3)
```



Normality of errors The QQ plot of residuals is used to check the normality assumptions. For this assumption to be satisfied, points on the normal probability plot show follow a straight line and and that can be observed from the plot below. Hence the residuals are normally distributed.

```
detach("package:car", unload=TRUE)

plot(model,2)
```



h. Fit a model using **sex**, and **exper** as explanatory variables. You might need to make an indicator variable for **sex**. Discuss your fit model, including an interpretation of all parameters in the model.

```
 banksal_n <- banksal %>%
mutate(sex = recode(sex,"MALE"= 1, "FEMALE" = 0))

model1 <- lm(bsal ~ sex+exper, data= banksal_n)
summary(model1)
```

```
##
## Call:
## lm(formula = bsal ~ sex + exper, data = banksal_n)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1224.42  -451.79    78.46   351.82  2202.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5016.2400   100.9967  49.667  < 2e-16 ***
## sex          814.0532   128.3907   6.340 8.95e-09 ***
## exper          1.2284     0.6743   1.822   0.0718 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 588.1 on 90 degrees of freedom
## Multiple R-squared:  0.328,  Adjusted R-squared:  0.313
## F-statistic: 21.96 on 2 and 90 DF,  p-value: 1.707e-08
```

The r square of 0.328 suggest that 32.8% of the data fits the regression model and since the r square is small , that means the model do not fits the observer data well. When the employee is a male, the intercept is 5830,

which is the average salary of a male given no experience and the experience coefficient of regression is 1.22 which represents the mean increases of beginners salary for every additional months of experience prior to starting work irrespective of the gender. Also the intercept for female is 5016 which means on average the female employee with no experience will receive a beginner salary of 5016.

h. What conclusions can be drawn about gender discrimination at Harris Trust based on your work above? Do these conclusions have to be qualified at all, or are they pretty clear cut?

From the analysis, we can conclude that there is a gender discrimination at Harris Trust. But the small value of r square 0.328 suggest that the data do not fit the model . Also using this sample data do not represent the population of interest that can be used to make a conclusion about the gender discrimination. We can add other factors and increase the sample size ,which could improve the significance of the fitted model and help us to make better conclusion of the gender discrimination.

i. Add an interaction term to the model in h. Give an interpretation of the new model and give a graphic representation of the data and fit model.

```
model2 <- lm(bsal ~ sex+exper+sex*exper, data= banksal_n)
summary(model2)
```

```
##
## Call:
## lm(formula = bsal ~ sex + exper + sex * exper, data = banksal_n)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1342.94  -534.40    39.31   369.75  2139.91
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4920.8166   115.2529  42.696  < 2e-16 ***
## sex         1042.8877   187.1620   5.572 2.65e-07 ***
## exper          2.1844     0.8804   2.481   0.0150 *
## sex:exper     -2.2507     1.3509  -1.666   0.0992 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 582.4 on 89 degrees of freedom
## Multiple R-squared:  0.3483, Adjusted R-squared:  0.3263
## F-statistic: 15.86 on 3 and 89 DF,  p-value: 2.426e-08
```

The intercept for male employees is 5963, which means on average male employees of not experience will take 5963 dollars as beginner salary and female employees with no experience will take 4921 dollars as beginner salary. The slope of the coefficient of experience would be -0.07 which means a month increase in experience of a male employee would decrease his salary by 0.07 dollars. Also the coefficient for experience of a female employee is 2.18 , which mean a month increase in experience would increase female salary by 2.18 dollars.

Still the model is have a lower r square value of 0.34 which mean 34% of the data is fits the model. Even though there is an improving from the previous model, there suggest be improvement but either changing that variables or adding more variables.

m. Often salary data is logged before analysis. Would you recommend logging starting salary in this study? Support your decision.

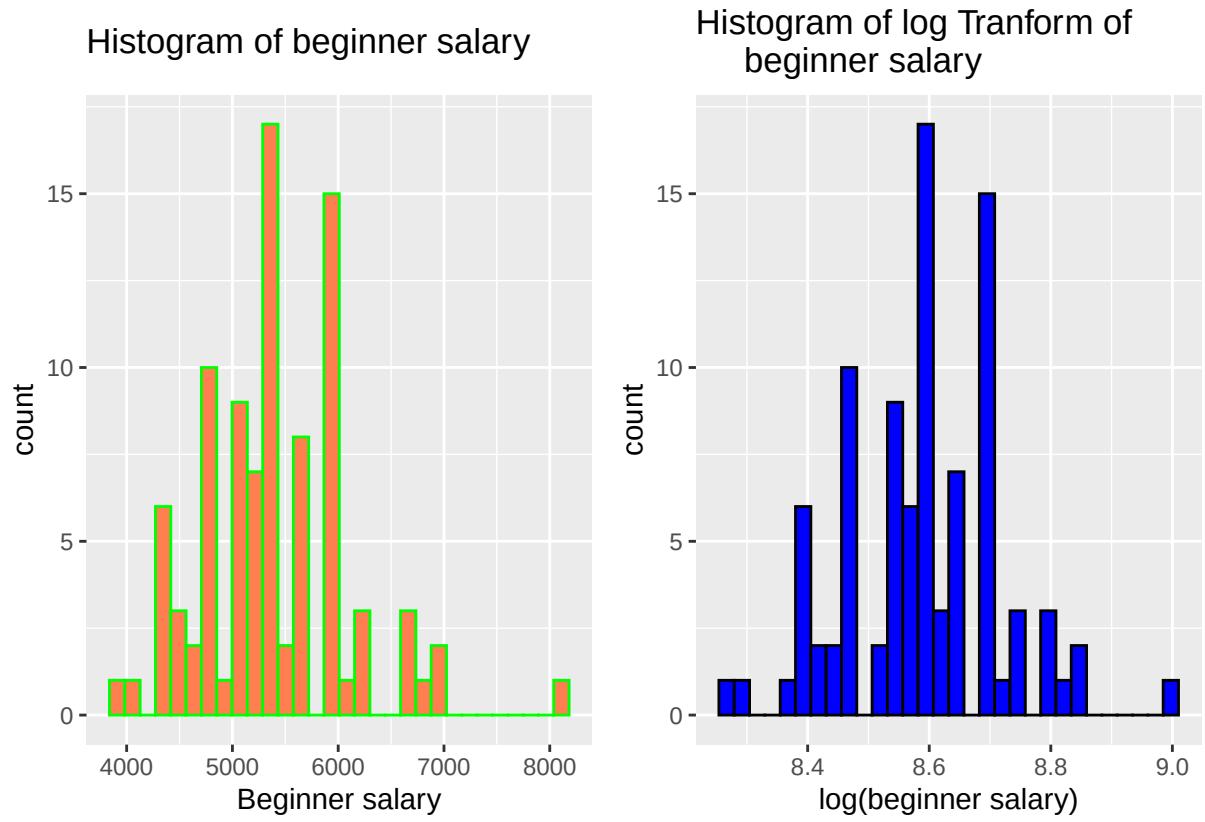
Salary is mostly recommended to be logged transfrom when performing analysis, but in this analysis, there is not need, since the data was normally distriuted and the graph below show it. Even after transformation,the graph is still the same.

```
new_bs <-banksal %>%
  mutate(bbsal =log(bsal))



plotb <- new_bs%>%
  ggplot(aes(x=bsal))+ geom_histogram(fill="coral",color="green")+
  labs(title="Histogram of beginner salary ",x="Beginner salary")

plotc <- new_bs %>%
  ggplot(aes(x= bbsal)) + geom_histogram(fill="blue",color="black")+
labs(title = "Histogram of log Tranform of
     beginner salary",x="log(beginner salary)")
plotb + plotc
```



Sitting and MTL thickness

Sitting and MTL thickness. @Siddarth2018 researched relations between time spent sitting (sedentary behavior) and the thickness of a participant's medial temporal lobe (MTL) in a 2018 paper entitled, "Sedentary behavior associated with reduced medial temporal lobe thickness in middle-aged and older adults". MTL volume is negatively associated with Alzheimer's disease and memory impairment. Their data on 35 adults can be found in `sitting.csv`. Key variables include:

- `MTL` = Medial temporal lobe thickness in mm
- `sitting` = Reported hours/day spent sitting
- `MET` = Reported metabolic equivalent unit minutes per week
- `age` = Age in years
- `sex` = Sex (`M` = Male, `F` = Female)

- `education` = Years of education completed

```
library(readr)
sitting <- read_csv("sitting.csv")
```

a. In their article's introduction, Siddarth et al. differentiate their analysis on sedentary behavior from an analysis on active behavior by citing evidence supporting the claim that, "one can be highly active yet still be sedentary for most of the day." Fit your own linear model with `MET` and `sitting` as your explanatory and response variables, respectively. Using R^2, how much of the variability in hours/day spent sitting can be explained by MET minutes per week? Does this support the claim that sedentary behaviors may be independent from physical activity?

```
model_n <- lm(sitting~ MET,data=sitting)
summary(model_n)
```

```
##
## Call:
## lm(formula = sitting ~ MET, data = sitting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6339 -2.3749 -0.3642  2.2418  8.0408
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.5015886  0.9145400   8.203  1.8e-09 ***
## MET         -0.0001982  0.0004708  -0.421    0.676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.365 on 33 degrees of freedom
## Multiple R-squared:  0.005345,   Adjusted R-squared:  -0.0248
## F-statistic: 0.1773 on 1 and 33 DF,  p-value: 0.6764
```

The squared is 0.0053 which means that 0.53% of the data fits the regression model taking, meaning 99.47% of variance of sitting can not be explained by MET. And since the p-value 0.67 for the F_statistic id greater than 0.05, when fail to reject than sedentary behaviors are independent from physical activity.

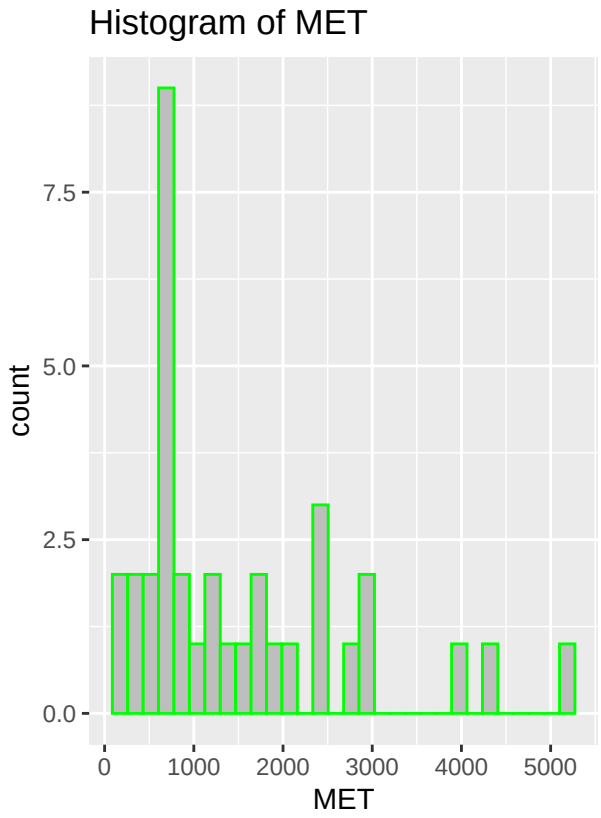
`b. In the paper's section, "Statistical analysis", the authors report that, "Due to the skewed distribu`

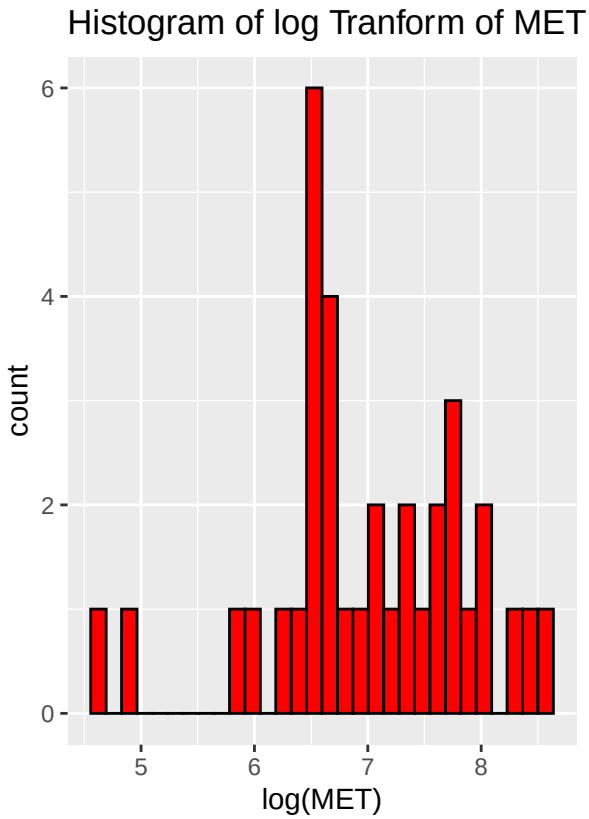
I agree with the papers decision to log-transform the MET, because the plot MET is right skewed which which would make interpreting very difficult.I used the colors to distinguish between the MET plot and the log-transformed MET plot.

```
sit <-sitting %>%
  mutate(MET_log = log(MET))

h_sit <- sit%>%
  ggplot(aes(x=MET))+ geom_histogram(fill="grey",color="green")+
  labs(title="Histogram of MET ",x="MET")

log_sit <- sit %>%
  ggplot(aes(x= MET_log)) + geom_histogram(fill="red",color="black")+
labs(title = "Histogram of log Tranform of MET",x="log(MET)")
h_sit + log_sit
```





c. Fit a model with `MTL` as the response and `sitting` as the sole explanatory variable. Are the lin

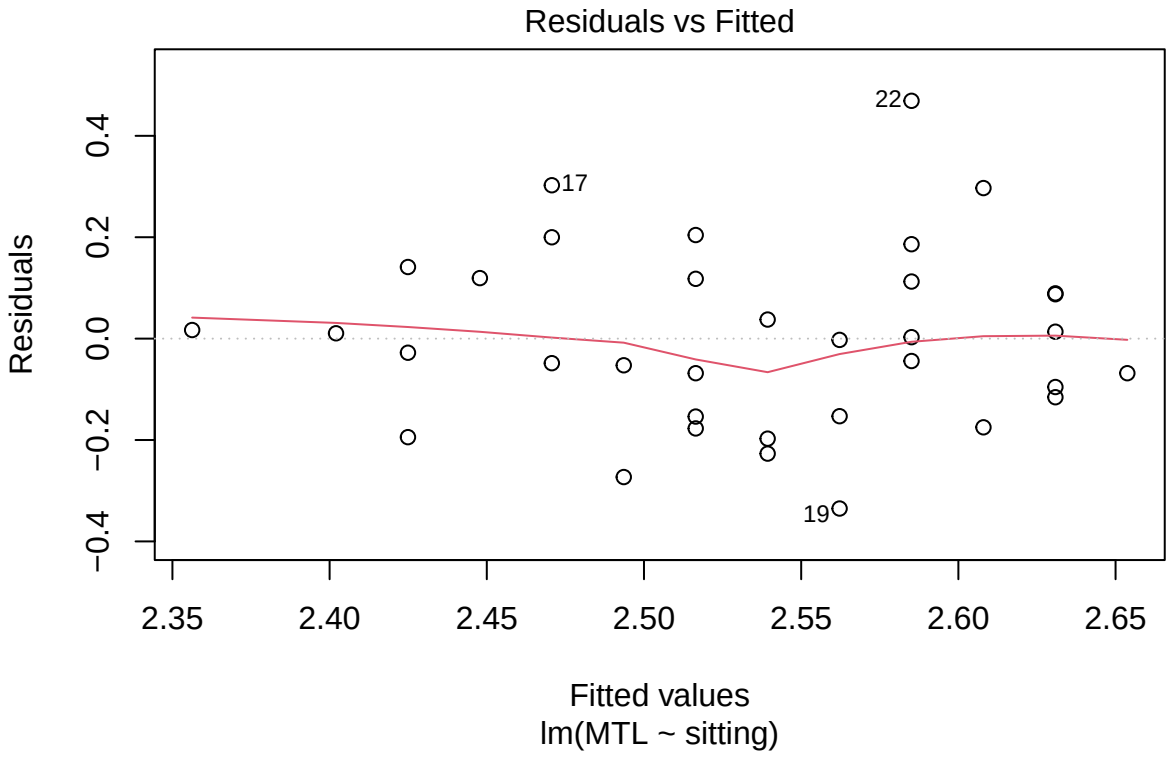
```
model_mtl <- lm (MTL ~ sitting, data=sitting)
summary(model_mtl)
```

```
##
## Call:
## lm(formula = MTL ~ sitting, data = sitting)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33511 -0.13432 -0.00252  0.11527  0.46907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.69951    0.07309  36.933   <2e-16 ***
## sitting     -0.02288    0.00924  -2.476   0.0186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1791 on 33 degrees of freedom
## Multiple R-squared:  0.1567, Adjusted R-squared:  0.1312
## F-statistic: 6.132 on 1 and 33 DF,  p-value: 0.01857
```

Validity : The variables MLT and sitting used to fit the model is valid and significant o the model. Representativenss: The sample data represents the population of interest.

Addivity and linearity: From the graph below , the red line is not straight along the zero residual, which means the relationship of MLT and mean of sitting is not linear.

```
plot(model_mtl,1)
```



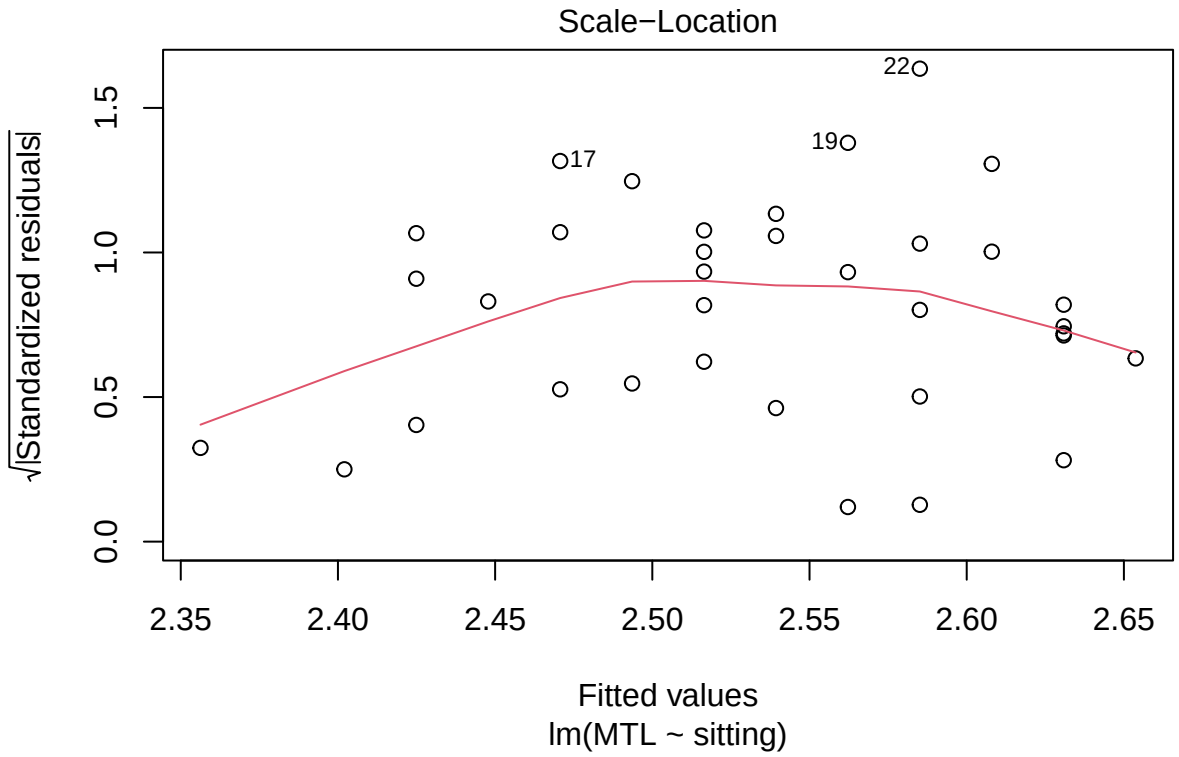
Independence of error: Using the durbin watson test, where p value 0.144 is greater than the alpha value of 0.05 we fail to reject the null hypothesis and concluded that errors are not autocorrelated and hence independent.

```
library(car)
durbinWatsonTest(model_mtl)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        0.207125      1.525642   0.158
##  Alternative hypothesis: rho != 0
```

Equal variance: From the plot there is an unequal spread of points and a curve red that which shows that the error are un equal.

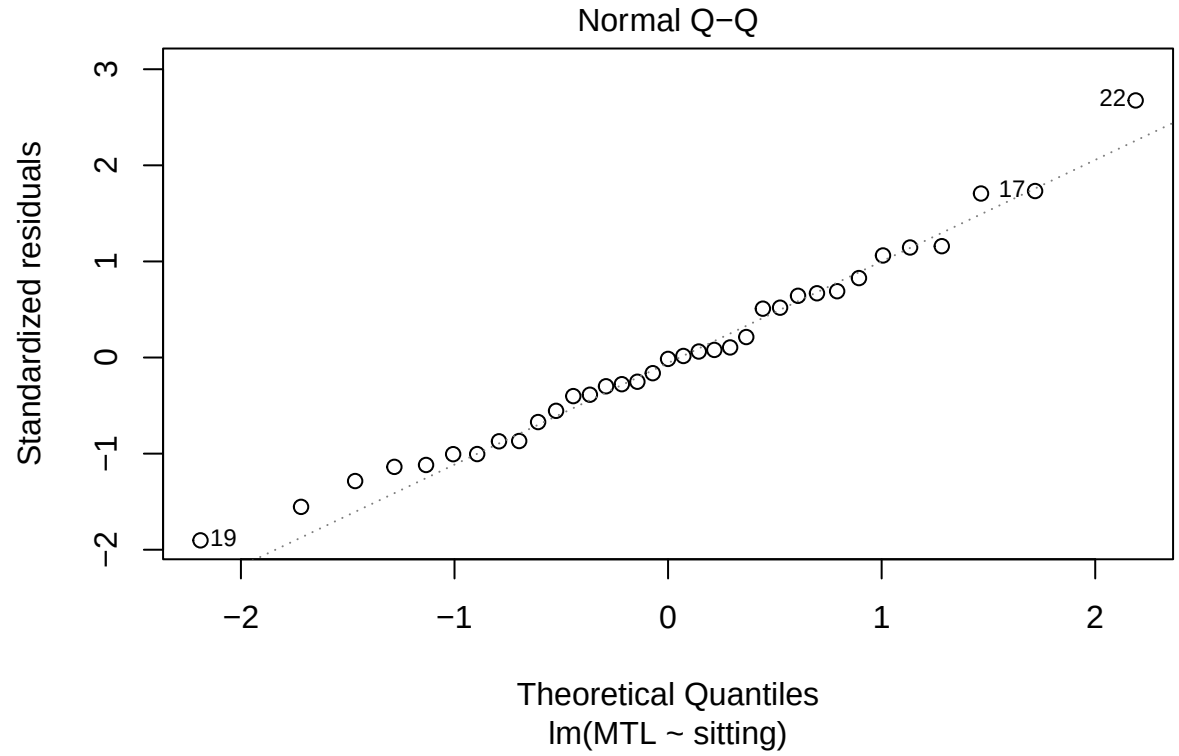
```
plot(model_mtl,3)
```



Normality of error:

From the QQplot , most of the points lie on the straight line and i can conclude that the errors are normally distributed.

```
plot(model_mtl,2)
```



d. One model fit in @Siddarth2018 includes `sitting`, log–transformed `MET`, and `age` as explanatory variables. Fit this model and get parameter estimates and SEs.

```
model_MET <- lm(MTL~ sitting+MET_log+age, data=sit)
summary(model_MET)
```

```
##
## Call:
## lm(formula = MTL ~ sitting + MET_log + age, data = sit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30673 -0.13091 -0.02548  0.13243  0.44786
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.363491   0.402649   5.870 1.78e-06 ***
## sitting     -0.020952   0.009500  -2.206    0.035 *
## MET_log      0.006911   0.035698   0.194    0.848
## age          0.004535   0.004022   1.127    0.268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1811 on 31 degrees of freedom
## Multiple R-squared:  0.1901, Adjusted R-squared:  0.1117
## F-statistic: 2.425 on 3 and 31 DF,  p-value: 0.08441
```

f. Based on your results from the previous part, do you support the paper's claim that, "it is possible that sedentary behavior is a more significant predictor of brain structure, specifically MTL thickness [than physical activity]"? Why or why not?

From the previous results, we found out that the p_value of the model is 0.01857 which is less than the alpha value of 0.05. Therefore we reject the null hypothesis that $\beta = O$, meaning there is a significant relationship between sedentary behavior (sitting) and the brain structure (MTL).

g. A *New York Times* article was published discussing @Siddarth2018 with the title "Standing Up at Your Desk Could Make You Smarter" [@Friedman2018]. Do you agree with this headline choice? Why or why not?

I do not agree with with this headline, because from the previous model, the r squared was 0.157 which mean 15% of of the data fits the model, that is about 85% of variance of MLT can not be explained by variance of sitting. This do not mean the model is wrong but other factors should ne consider as such sample size, other variables that could have effect to the MLT.

Simulating two normal distribution

Simulate two Normally distributed random variables X1 and X2 with correlation 0.8, both should have a mean value of 70 and a standard deviation of 8. See https://www.probabilitycourse.com/chapter5/5_3_2_bivariate_normal_dist.php equation 5.23 for formulas that will help you see how to do the simulation. There are other ways to do this as well. Think of X2 as a score on the second exam in a class, and X1 as the score on the first exam.

a. Analytically find the expected value and variance of `Change = X2 - X1`.
The expected of X2-X1 is $E(X2 - X1) = E(X2) - E(X1) = 70 - 70 = 0$ and the variance is $var(X2 - X1) = \sigma_2^2 + \sigma_1^2 - 2\rho\sigma_1\sigma_2 = 25.6$ and the computation is shown below.

$$(X_2 - X_1) \sim N(0, 25.6)$$

```
mu1= 70
mu2=70
sigma1=8
sigma2=8
rho = 0.8

E_means = mu2-mu1
E_means
```

```
## [1] 0
```

```
var_x2x1 = sigma1^2+sigma1^2-2*rho*sigma1*sigma2
var_x2x1
```

```
## [1] 25.6
```

b. Use a simulation with n=1000 to find the mean and variance of `Change` using simulation. The expected value for change is 0.4054 and variance 25.4487

```
x_1= rnorm(1000,mean=70,8)
x_2= rnorm(1000,mean=70,8)

change= x_2-x_1

m1 = mean(x_2)-mean(x_1)
m1
```

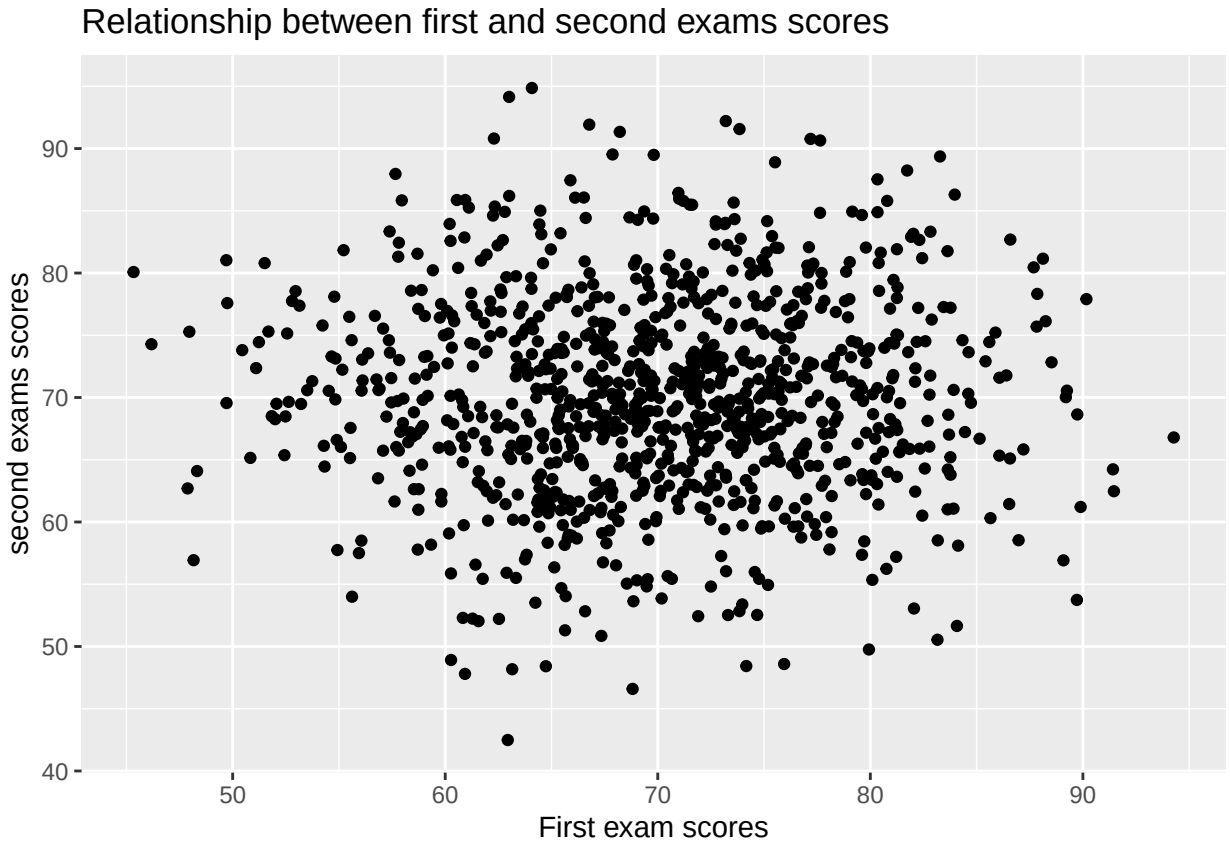
```
## [1] 0.02123821
```

```
vars = var(x_2)+var(x_1)-2*rho*sd(x_1)*sd(x_2)
vars
```

```
## [1] 26.40931
```

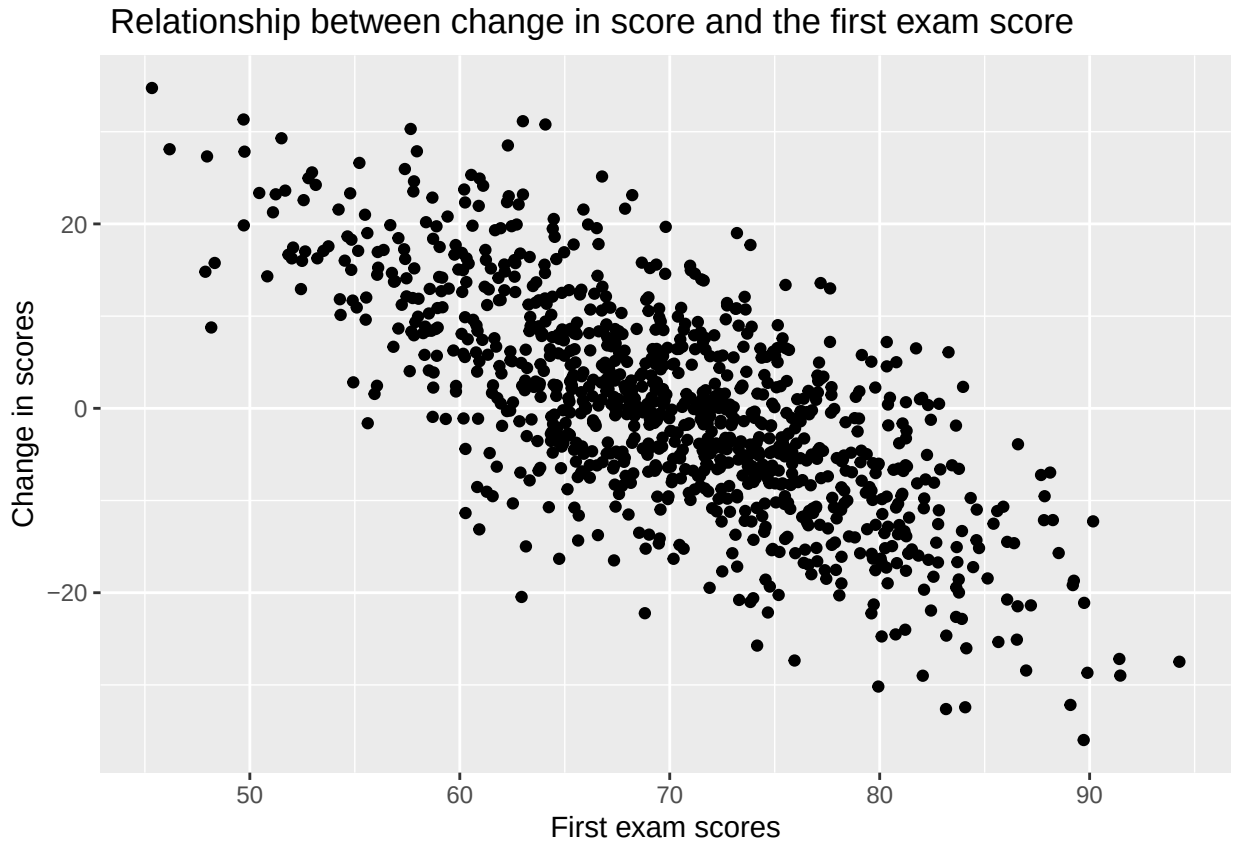
c. Plot your simulated data on a scatterplot.

```
  data_sim <- as.tibble(x_1,x_2,change)
data_sim%>%
  ggplot(aes(x_1,x_2))+geom_point()+
  labs(title = "Relationship between first and second exams scores",
       x="First exam scores",y="second exams scores")
```



d. Also plot X1 vs. Change.

```
data_sim%>%
  ggplot(aes(x_1,change))+geom_point()+
  labs(title = " Relationship between change in score and the first exam score",
       x="First exam scores",y="Change in scores")
```



e. How might regression to the mean cause issues in assessing whether or not a student improved or did worse on the second exam compared to the first?

Regression to the mean would cause issue for student to perform worse in the second exams since less and high extreme score would regress towards the mean which in respective would reduce tha score of the second exam. The graph above shows a negative relationship between the change in score and the first exam scores. Simply , regression of the mean is when extreme value from first sampling is picked out again in the second sampling which is seen getting closer to the previous mean.