# CSCI E-106:Assignment 1

**Due Date: September 22, 2023 at 11:59 pm EST**

**Instructions**

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

---

## Problem 1

Refer to the Grade point average Data. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). (40 points)

a-) Import the data into r (10 points)

b-) Plot the ACT against GPA and comment on the relationship (10 points)

c-) Calculate the correlation between ACT and GPA (10 points)

d-) Build a regression model and comment on the intercept and slope (10 points)

```
# Load the data into a data frame
gpadata <- read.csv("Grade Point Average Data.csv")

# Check the structure of the data
str(gpadata)
```

```
## 'data.frame':    120 obs. of  2 variables:
##  $ Y: num  3.9 3.88 3.78 2.54 3.03 ...
##  $ X: int  21 14 28 22 21 31 32 27 29 26 ...
```

```
# Print the first 10 rows of the data
head(gpadata, 10)
```

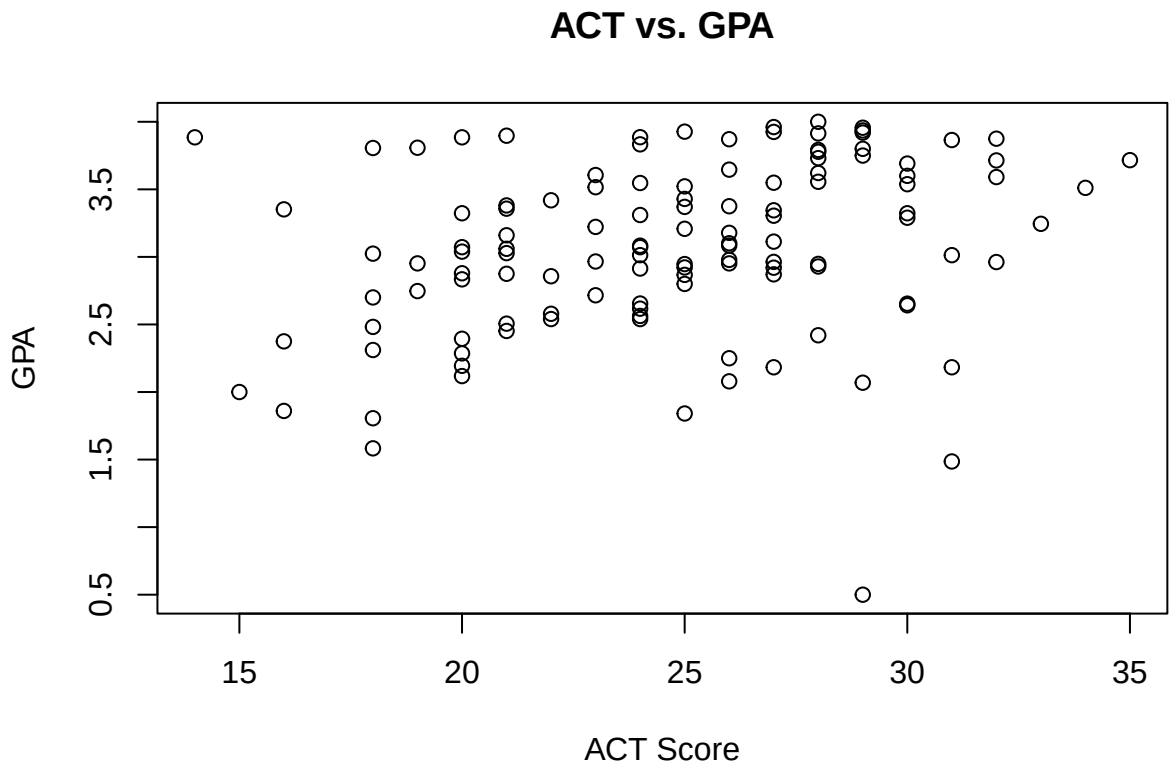
```
##        Y  X
## 1  3.897 21
## 2  3.885 14
## 3  3.778 28
## 4  2.540 22
```

```
## 5  3.028 21
## 6  3.865 31
## 7  2.962 32
## 8  3.961 27
## 9  0.500 29
## 10 3.178 26
```

The above code imports the Grade Point Average Dataset, shows its structure and prints the first 10 rows. We observe that there are 2 variables, ACT score and GPA, each having 120 observations.

```r
# Assign X to ACT scores and Y to GPA
ACT <- gpadata$X
GPA <- gpadata$Y

# Plot ACT against GPA
plot(ACT, GPA, main = "ACT vs. GPA", xlab = "ACT Score", ylab = "GPA")
```




The above code plots the ACT score against GPA. We observe a slightly positive relationship although there are some outliers i.e as the ACT score increases, the GPA is predicted to be higher.

```r
# Calculate the correlation coefficient
correlation <- cor(ACT, GPA)
print(correlation)
```

```
## [1] 0.2694818
```

We observe that there is a low positive correlation of 0.27 between the ACT score and the outcome, GPA. Further insight can be obtained by building a regression model.

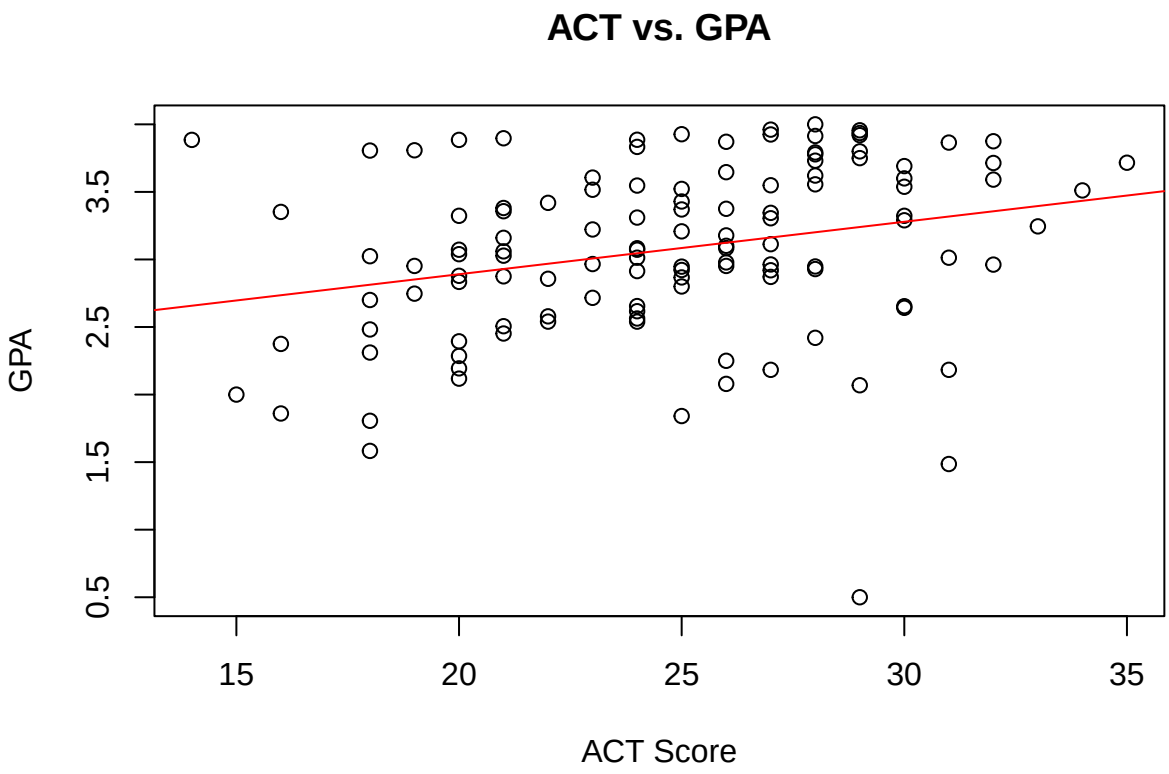
```r
# Build a regression model
model <- lm(GPA ~ ACT, data = gpadata)

# Print the model summary
summary(model)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = gpadata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## ACT          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

```r
# Plot ACT against GPA
plot(ACT, GPA, main = "ACT vs. GPA", xlab = "ACT Score", ylab = "GPA")

# Add a red regression line
abline(model, col = "red")
```




From the linear regression model, we see that the the regression line is Y = 2.11405 + 0.03883X, which confirms the earlier hypothesis from the plot that GPA increases as the ACT score increases. The slope of the regression line is however small, which tells us that GPA does not increase significantly as ACT score increases. The adjusted R-square value of 6.47% confirms the weak relationship between GPA and ACT score.

## Problem 2

The dataset uswages is drawn as a sample from the Current Population Survey in 1988. You can download this data set by installing faraway library. To get the data set, copy and paste the r command: install.packages("faraway"); data(uswages, package="faraway").

The wage is the response variable. Please see below for the full list of variables.

wage: Real weekly wages in dollars (deflated by personal consumption expenditures - 1992 base year)

educ:Years of education

exper:Years of experience

race:1 if Black, 0 if White (other races not in sample)

smsa:1 if living in Standard Metropolitan Statistical Area, 0 if not

ne:1 if living in the North East

mw:1 if living in the Midwest

we:1 if living in the West

so:1 if living in the South

pt:1 if working part time, 0 if not

a-) How many observations are in the data set?

b-) Calculate the mean and median of each variable? Are there any outliers in the data set?

c-) Calculate the correlation among wage,education and experience. Plot each of the predictors against the response variable. Identify the variables that are strongly correlated with the response variable.

d-) Is there difference in wages based on race?

e-) Build a regression model by using only education to predict the response variable. State the regression model.

f-) Build a regression model by using only experience to predict the response variable. State the regression model.

```
# Set a specific CRAN mirror as code was showing error while knitting
options(repos = c(CRAN = "https://cran.rstudio.com/"))

# Install and load the "faraway" library
install.packages("faraway")
```

```
##
## The downloaded binary packages are in
##  /var/folders/06/tcq9vzm15837hsvfb646g0gh0000gp/T//Rtmph5ATa3/downloaded_packages
```

```
library(faraway)

# Load the "uswages" dataset
data(uswages, package = "faraway")
```

```
# Show the structure of the dataset
str(uswages)
```

```
## 'data.frame':    2000 obs. of  10 variables:
##  $ wage : num  772 617 958 617 902 ...
##  $ educ : int  18 15 16 12 14 12 16 16 12 12 ...
##  $ exper: int  18 20 9 24 12 33 42 0 36 37 ...
##  $ race : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ smsa : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ ne   : int  1 0 0 1 0 0 0 0 0 0 ...
##  $ mw   : int  0 0 0 0 1 0 0 1 0 1 ...
##  $ so   : int  0 0 1 0 0 0 1 0 0 0 ...
##  $ we   : int  0 1 0 0 0 1 0 0 1 0 ...
##  $ pt   : int  0 0 0 0 0 0 1 1 1 0 ...
```

```
# Number of observations
num_observations <- nrow(uswages)
cat("Number of observations:", num_observations, "\n")
```

```
## Number of observations: 2000
```

```r
# Print the first 10 rows of the dataset
head(uswages, 10)
```

```
##         wage educ exper race smsa ne mw so we pt
## 6085  771.60   18    18    0    1  1  0  0  0  0
## 23701 617.28   15    20    0    1  0  0  0  1  0
## 16208 957.83   16     9    0    1  0  0  1  0  0
## 2720  617.28   12    24    0    1  1  0  0  0  0
## 9723  902.18   14    12    0    1  0  1  0  0  0
## 22239 299.15   12    33    0    1  0  0  0  1  0
## 14379 541.31   16    42    0    1  0  0  1  0  1
## 12878 148.39   16     0    0    1  0  1  0  0  1
## 23121 273.19   12    36    0    1  0  0  0  1  1
## 13086 666.67   12    37    0    0  0  1  0  0  0
```

There are 2000 observations are in the data set, 9 of which are of the integer datatype and 1 is of num/float datatype.

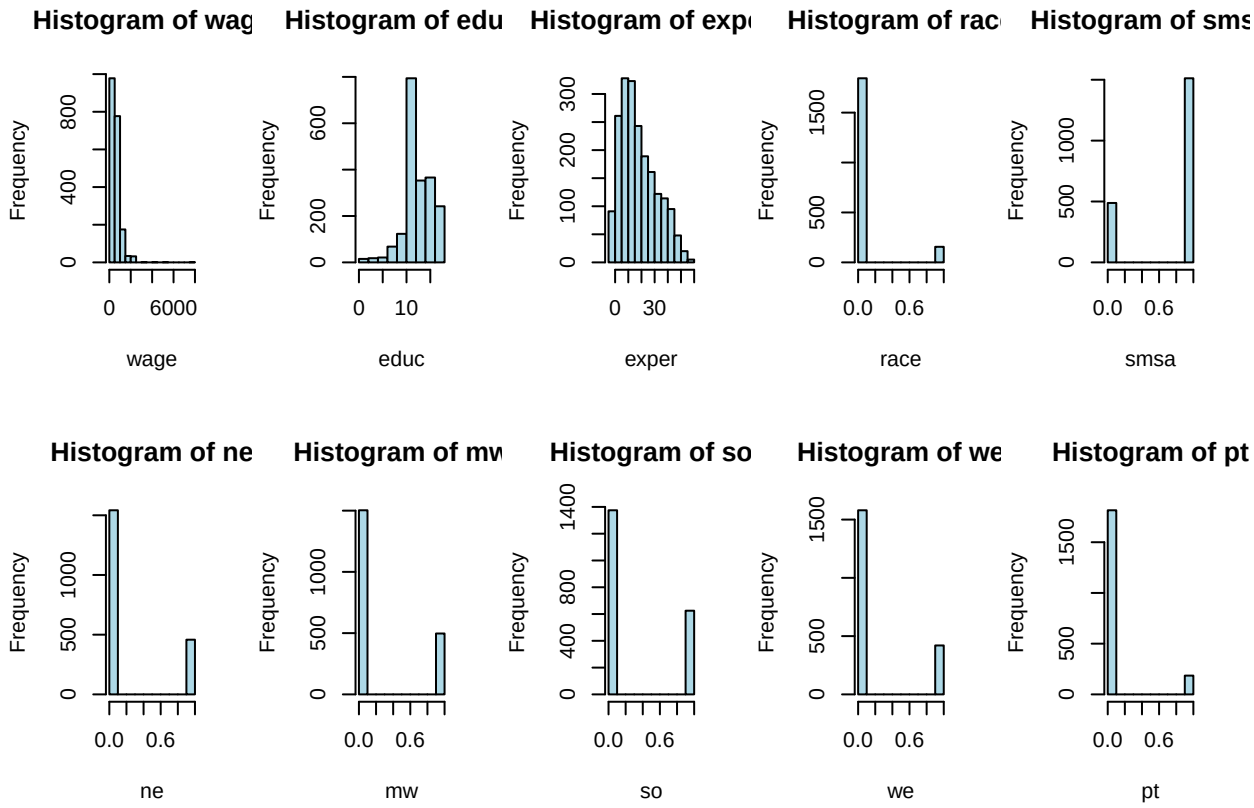
```r
# Summary statistics for all variables
summary(uswages)
```

```
##       wage              educ           exper            race
##  Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000
##  1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000
##  Median : 522.32   Median :12.00   Median :15.00   Median :0.000
##  Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000
##       smsa             ne              mw              so
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##        we              pt
##  Min.   :0.00   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000
##  Median :0.00   Median :0.0000
##  Mean   :0.21   Mean   :0.0925
##  3rd Qu.:0.00   3rd Qu.:0.0000
##  Max.   :1.00   Max.   :1.0000
```

```r
# List of variables for which histograms will be created
variables <- c("wage", "educ", "exper", "race", "smsa", "ne", "mw", "so", "we", "pt")

# Set up a 2x5 grid for plotting histograms
par(mfrow = c(2, 5))

# Create histograms for each variable
for (variable in variables) {
  hist(uswages[, variable], main = paste("Histogram of", variable), xlab = variable, ylab = "Frequency"
}
```




```
# Reset the plotting layout
par(mfrow = c(1, 1))
```

In the above code, we have displayed the summary of the "uswages" dataset as well as plotted histograms for each variable to check for outliers. We see that the mean and median values differ for the variables "wage", "educ" and "exper", and they are almost similar for the variables "race", "smsa", "ne", "so", "we" and "pt".

We know that outliers cause a skewed distribution resulting in a larger difference between the mean and median. Hence we plotted histograms, from which we can comment that there are outliers in the dataset. The histogram is positively skewed for the variables "wage", "educ", and "exper", which means that the mean is larger than the median.

```
# Calculate the correlation between Education and Wage
correlation1<- cor(uswages$educ, uswages$wage)
print(correlation1)
```

```
## [1] 0.2483358
```

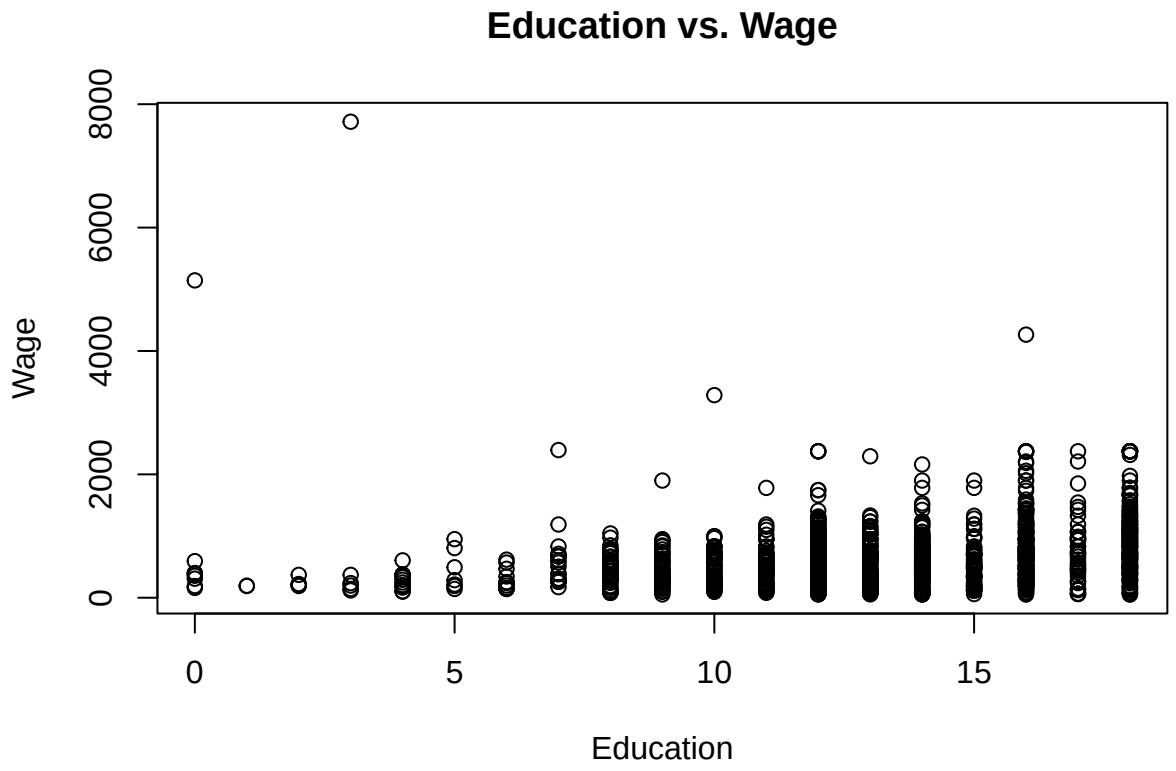
```
# Calculate the correlation between Experience and Wage
correlation2<- cor(uswages$exper, uswages$wage)
print(correlation2)
```

```
## [1] 0.1832012
```

```r
# Calculate the correlation between Experience and Education
correlation3<- cor(uswages$exper, uswages$educ)
print(correlation3)
```

```
## [1] -0.3024788
```

```r
# Plot Experience and Education against the response variable "wage"
plot(uswages$educ, uswages$wage, main = "Education vs. Wage", xlab = "Education", ylab = "Wage")
```




```r
plot(uswages$exper, uswages$wage, main = "Experience vs. Wage", xlab = "Experience", ylab = "Wage")
```


We observe that there is: - a low positive correlation of 0.25 between Education and Wage - a low positive correlation of 0.18 between Experience and Wage - a moderate negative correlation of -0.30 between predictor variables Experience and Education, i.e there may be multicollinearity

We have also plotted Experience and Education against the response variable Wage. From the plots, we observe that Education has a slightly strong correlated with Wage, and Experience has a slightly weaker correlation with Wage comparatively

```r
# Separate wages by race
wage_black <- uswages$wage[uswages$race == 1]
wage_white <- uswages$wage[uswages$race == 0]

# Perform a t-test
t_test_result <- t.test(wage_black, wage_white)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  wage_black and wage_white
## t = -6.1253, df = 221.05, p-value = 4.096e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -218.0177 -111.8771
## sample estimates:
## mean of x mean of y
##  456.0363  620.9838
```

To figure out if there difference in wages based on race, we have conducted a two sample T-test between "wage_black" and "wage_white". - We observe that the p-value is $< 0.05$ and extremely small. This means that there is strong statistical evidence to reject the null hypothesis that the means of the two groups are equal. - The negative t-value and the 95% confidence interval both indicate that the "wage_black" group tends to have significantly lower wages compared to the "wage_white" group.

```
# Build a regression model with education
model_education <- lm(wage ~ educ, data = uswages)
summary(model_education)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = uswages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -743.6 -269.5  -67.7  173.0 7492.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  109.754     44.616    2.46    0.014 *
## educ          38.011      3.317   11.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 445.5 on 1998 degrees of freedom
## Multiple R-squared:  0.06167,    Adjusted R-squared:  0.0612
## F-statistic: 131.3 on 1 and 1998 DF,  p-value: < 2.2e-16
```

```
# Plot Education against the response variable "wage"
plot(uswages$educ, uswages$wage, main = "Education vs. Wage", xlab = "Education", ylab = "Wage")

# Add a red regression line
abline(model_education, col = "red")
```


The regression model has been built above and regression line plotted. The regression model can be stated as: Weekly Wage = 109.764 + 38.011 (Years of Education)

The slope of the regression line is however small, which tells us that "wage" does not increase significantly as "educ" increases. The adjusted R-square value of 6.12% confirms the weak relationship between "wage" and "educ".

```
# Build a regression model with experience
model_experience <- lm(wage ~ exper, data = uswages)
summary(model_experience)
```

```
##
## Call:
## lm(formula = wage ~ exper, data = uswages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -755.5 -271.0  -77.5  165.7 6852.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 492.1669    17.2043   28.61   <2e-16 ***
## exper         6.2981     0.7561    8.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 452.2 on 1998 degrees of freedom
## Multiple R-squared:  0.03356,    Adjusted R-squared:  0.03308
## F-statistic: 69.39 on 1 and 1998 DF,  p-value: < 2.2e-16
```

```r
# Plot Experience against the response variable "wage"
plot(uswages$exper, uswages$wage, main = "Experience vs. Wage", xlab = "Experience", ylab = "Wage")

# Add a red regression line
abline(model_experience, col = "red")
```


The regression model has been built above and regression line plotted. The regression model can be stated as: Weekly Wage = 492.1669 + 6.2981 (Years of Experience)

The slope of the regression line is even smaller than that for Education, which tells us that "wage" does not increase significantly as "exper" increases. The adjusted R-square value of 3.31% confirms the weaker relationship between "wage" and "exper" than that between "wage" and "educ".