# LInear Regression

*Vazgen Tadevoysan*

*February 13, 2019*

For this Homework, you are required to submit both Markdown and HTML files with your answers and code in it. Be sure that the .Rmd file is working, so when I run it, there would be no errors and represent the same information as HTML. Write your code and interpretations under each question. The interpretations of the results need to be written below or above all the charts, summaries, or tables. Do not remove problems from your Markdown file.

Use gpafactors.csv dataset uploaded on Moodle to analyze the relationship between grade point average of students and different factors. The description of the variables is given in a separate file. Pay close attention to the names of axes, titles, and labels.

Problem 1. 1 pt.

Load the file.

Get rid of variables that are irrelevant for regression analysis using function select().

Check whether the data types are correct, if not make appropriate corrections assigning labels to each level according to the data description.

```r
library(plyr)
library(ggplot2)
library(gridExtra)
library(ggcorrplot)



df<-read.csv("gpafactors.csv",row.names = 1)
str(df)
df[duplicated(df$studentid),][1]
sapply(df, function(x) sum(is.na(x)))
str(df)
df$Months <- factor(months(as.Date(df$surveydate)))
df$imp<-ordered(df$imp,levels=c(1,2,3,4,5),labels=c('Not important','Apathetic','Neutral','Important','
df$job<-revalue(df$job, c("empl"="Employed", "unempl"="Uneployed"))
Data<-subset(df,select = -c(studentid,surveydate))
table(Data$Months)
```

Before data analysis and modeling it is required to do several steps of Data preprocessing. In the code above R reads csv file skipping first column containing index values. Furthermore it checks whether the dataset contains duplicated rows or missing values. Additionally, new “Data” dataframe is subsetted from the original excluding unnecessary variables after some data correction is done. The type of the column surveydate was factor, so it was necessary to make it as date then months was extracted from it. The action was done due to the fact there were 158 levels (158 days) and I preferred to have only months as factor decreasing number of levels to 6. I wanted to identify whether people lose their interest in studying after some month and or semester and as two variables that describe it (Month,imp) are categorical I can only count them

```r
new<-table(Data$Months,Data$imp)
print(new)
```

```
##
##            Not important Apathetic Neutral Important Very important
##   April               84        65      86        79             64
##   February              62          77       70          65                85
##   January               78          77       76          73                91
##   June                  21          17       17          17                15
##   March                 71          84       72          77                87
##   May                   82          84       75          74                75
```

```
chisq.test(new)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  new
## X-squared = 16.268, df = 20, p-value = 0.6999
```

```
Data$Months<-NULL
```

In the beginning of the year Gpa was very important for big part of surveyed people and Very Important level was the most frequent among them but later it is not anymore that frequent among people surveyed, for ex in April or in June. However portions of importance levels per group of months are not seemed to be very different from each other so I checked it using Pearson's Chi-squared Test for Count Data p-value is greater than alpha so I fail to reject H0 and can claim that month do not affect importance of gpa for student. Consequently I drop month column.

Problem 2. 2 pt.

Find the two most correlated numeric variables with grade point average of students using cor() and pairs() functions. Comment on it.

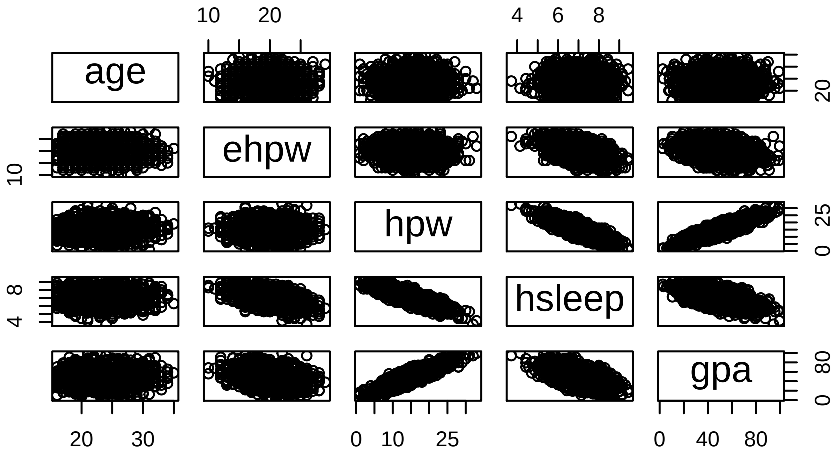
Find the binary variables which affect grade point average of students using boxplots. Comment on it.

```
only_num<-dplyr::select_if(Data, is.numeric)
cr <- cor(only_num)
# This is to remove redundancy as upper correlation matrix == lower
cr[upper.tri(cr, diag=TRUE)] <- NA
cr<-reshape2::melt(cr, na.rm=TRUE, value.name="cor")
cr<-cr[cr$Var1=="gpa",]
rownames(cr) <- NULL
cr
```

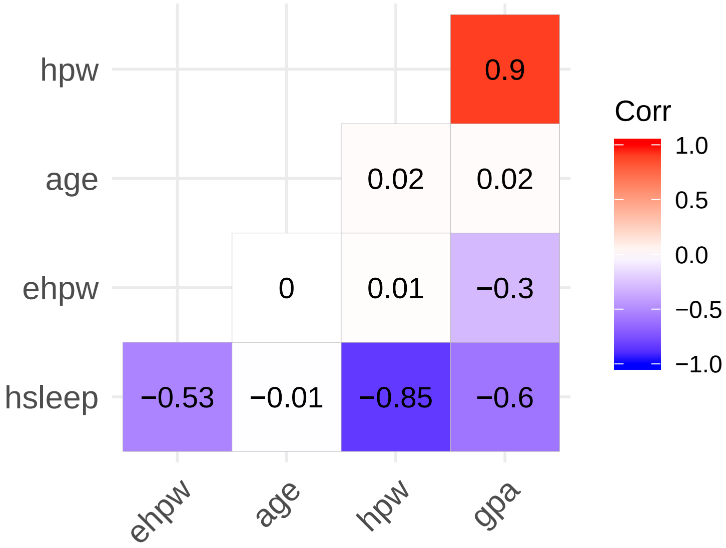
```
##   Var1   Var2         cor
## 1  gpa    age  0.01667937
## 2  gpa   ehpw -0.30058258
## 3  gpa    hpw  0.89829278
## 4  gpa hsleep -0.60251342
```

The two most correlated numeric variables with grade point average of student are variables Hours spent on studying a week and Hours of sleep per day.

```
pairs(only_num)
```

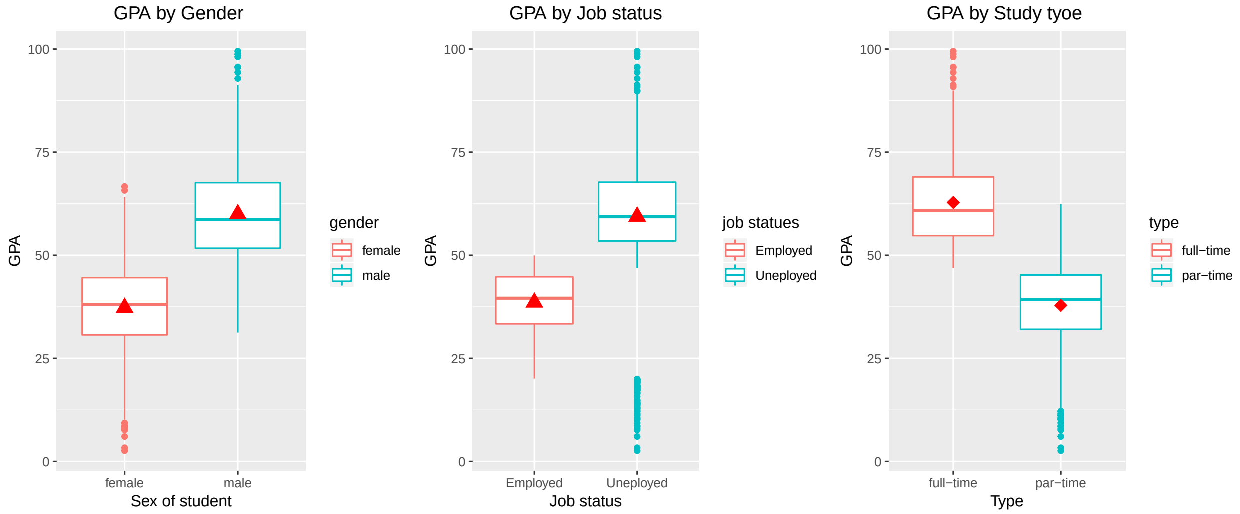


```
ggcorrplot(cor(only_num), hc.order = TRUE, type = "lower",
   lab = TRUE)
```



Plot provided by pairs() function states the same statement. Variables that are the most correlated with Gpa are hpw and hsleep , however they seemed to be correlated with each other. Hence ggcorplot() was used to get all correlations to make assumption about multycolineary for further modeling.

```
g1 = ggplot(data = Data, aes(y = gpa, x = gender,color = gender))+
  geom_boxplot()+
  scale_x_discrete(labels= levels(Data$gender) )+
  stat_summary(fun.y = "mean",geom = 'point',col='red', shape=17, size=4)+
  labs(title ="GPA by Gender " ,color =  "gender",y='GPA',x="Sex of student")+
  theme(plot.title = element_text(hjust = 0.5))
g2 = ggplot(data = Data, aes(y = gpa, x = job ,color = job ))+
  geom_boxplot()+
  scale_x_discrete(labels= levels(Data$job) )+
  stat_summary(fun.y = "mean",geom = 'point',col='red', shape=17, size=4)+
  labs(title ="GPA by Job status" ,color =  "job statues",y='GPA',x="Job status")+
  theme(plot.title = element_text(hjust = 0.5))
g3 = ggplot(data = Data, aes(y = gpa, x = type,color = type))+
  geom_boxplot()+
  stat_summary(fun.y = "mean",geom = 'point',col='red', shape=18, size=4)+
  labs(title ="GPA by Study tyoe " ,y='GPA',x="Type")+
  theme(plot.title = element_text(hjust = 0.5))
grid.arrange(g1, g2, g3,nrow=1)
```



Plot of boxplot derives some assumption which will be checked in the model. First plot states that sex of respondent affect GPA,males are more likely to get high GPA than women. Additionally, employed people's gpa more often is less than gpa-s of unemployed people. Furthermore, students who study part time tend to have lower gpa than full time students.

Problem 3. 4 pt.

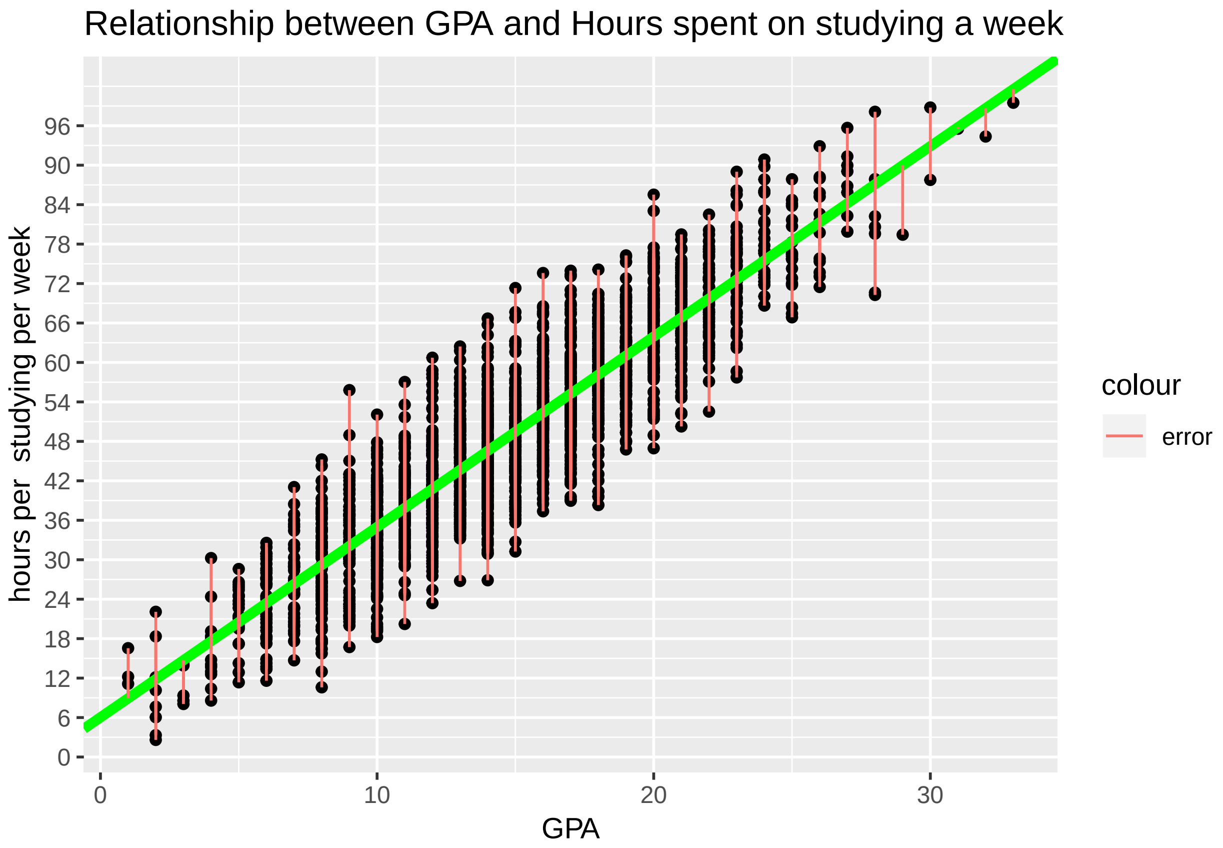
Use the lm() function to perform a simple linear regression with GPA as the response and the most correlated numeric variable as the predictor. Use the summary() function to print the results. Comment on the output:

a. Explain the meanings of coefficients (do they all have a meaning)?
b. Why do we need to add b0 to the regression model?
c. Which coefficients are significant (in which level)? Why and why not?
d. Explain the meaning of R-squared in your model.
e. Plot the response and the predictor. Use the geom_abline() function to display the least squares regression line.

```
model<-lm(gpa~hpw,data = Data)
summary(model)
```

```
##
## Call:
## lm(formula = gpa ~ hpw, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8197  -4.9597   0.0725   4.7562  23.7033
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.03372    0.49801   12.12   <2e-16 ***
## hpw          2.89477    0.03168   91.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.95 on 1998 degrees of freedom
## Multiple R-squared:  0.8069, Adjusted R-squared:  0.8068
## F-statistic:  8351 on 1 and 1998 DF,  p-value: < 2.2e-16
```

```
ggplot(Data,aes(x = hpw,y=gpa))+geom_point()+geom_abline(intercept =6.03372,slope = 2.89477,size=1.8,col
  labs(title = "Relationship between GPA and Hours spent on studying a week",x="GPA",y='hours per  stud
  scale_y_continuous(breaks=seq(0,100,6))+geom_segment(aes(x=hpw, xend=hpw, y=gpa, yend=model$fitted.va
```



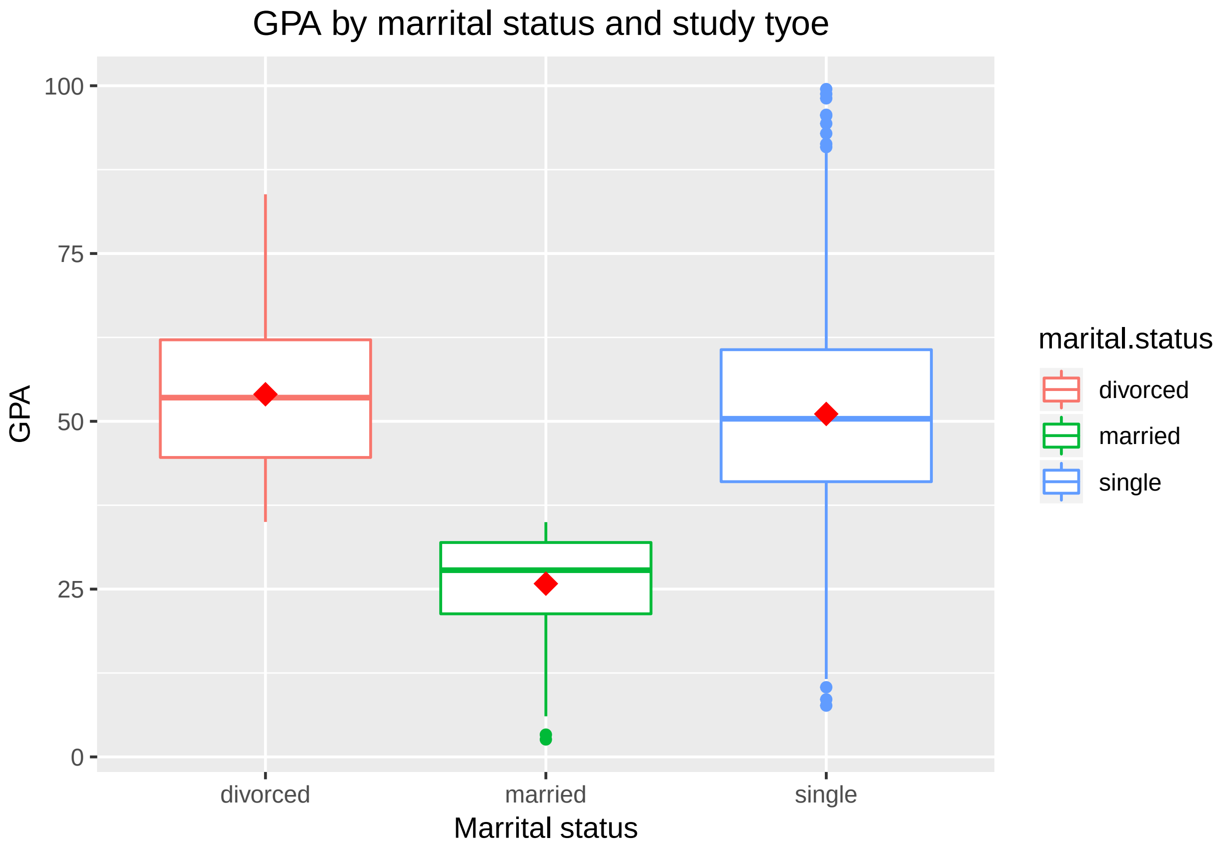
Our H0 is that coefficient of a variable is equal to 0, H1 it is different. Before conducting test we should define significance level(in our case I set it a to 0.05 ), also denoted as alpha or a, is the probability of rejecting the null hypothesis when it is true. P-values were less than alpha so we reject H0 and we claim that variables are significant for GPA. B0 is the intercept of the regression line with the y-axis. In other words it is the value of Y if the values of Xs = 0 .However, this is not a meaningful interpretation for this model because both X1 and X2 cannot be 0,so B0 just helps to set the regression line in the right place. R-squared-Hours spent on studying a week explains 80.6% of the variance of GPA.

Problem 4. 6 pt.

a. Discover the relationship between marital status and grade point average of the students using boxplot.
b. Run the regression model with GPA as a response and marital status as explanatory variable.
   In this regression model, the reference group for the categorical variable should be the single value. Use relevel().
c. Interpret the coefficients of a categorical variable.
d. Add to previous model one numeric variable. Do not apply summary(). Plot the response and the

predictors. You must have 3 labeled regression lines.

```
ggplot(data = Data, aes(y = gpa, x = marital.status,color = marital.status))+
  geom_boxplot()+scale_x_discrete(labels= levels(Data$marital.status))+
  stat_summary(fun.y = "mean",geom = 'point',col='red', shape=18, size=4)+
  labs(title ="GPA by marrital status and study tyoe " ,y='GPA',x="Marrital status")+
  theme(plot.title = element_text(hjust = 0.5))
```

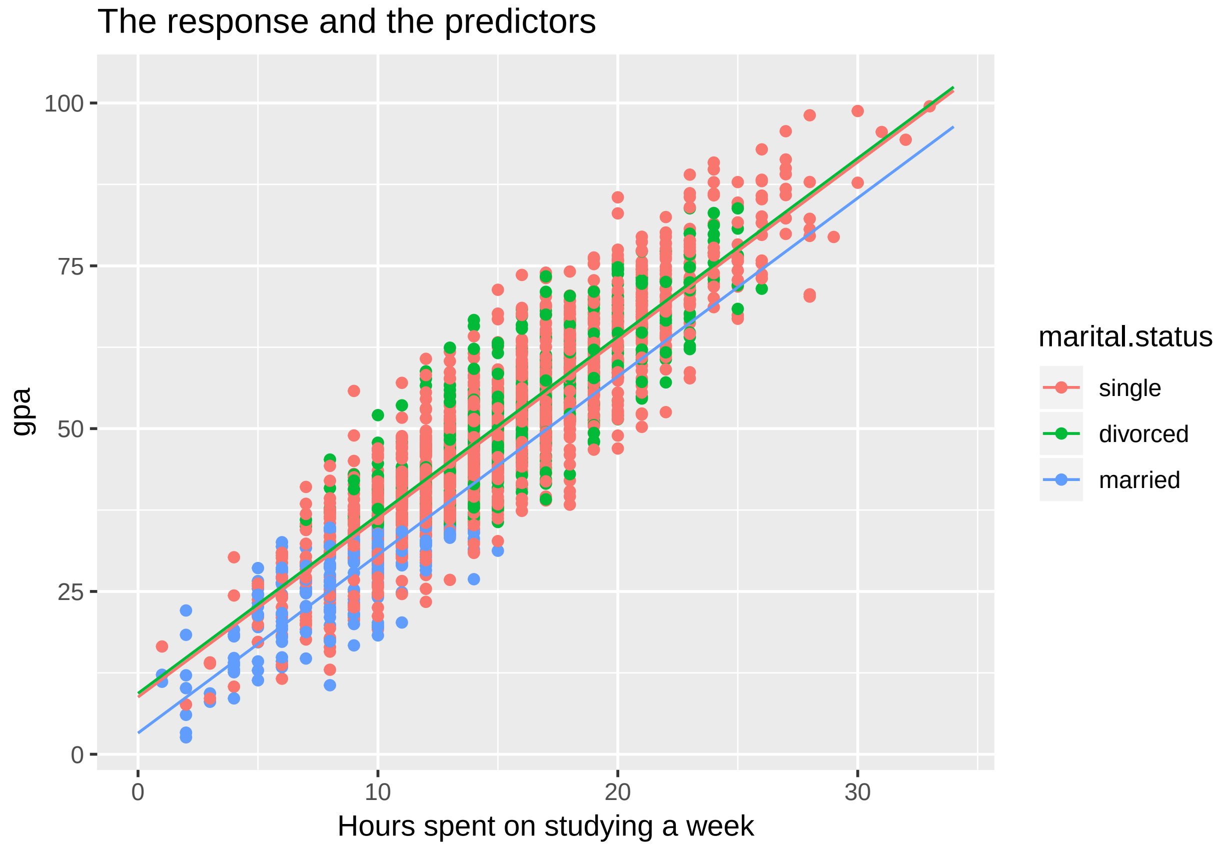


```
Data$marital.status<-relevel(Data$marital.status,ref = 'single')
model2<-lm(gpa~marital.status,data = Data)
summary(model2)
```

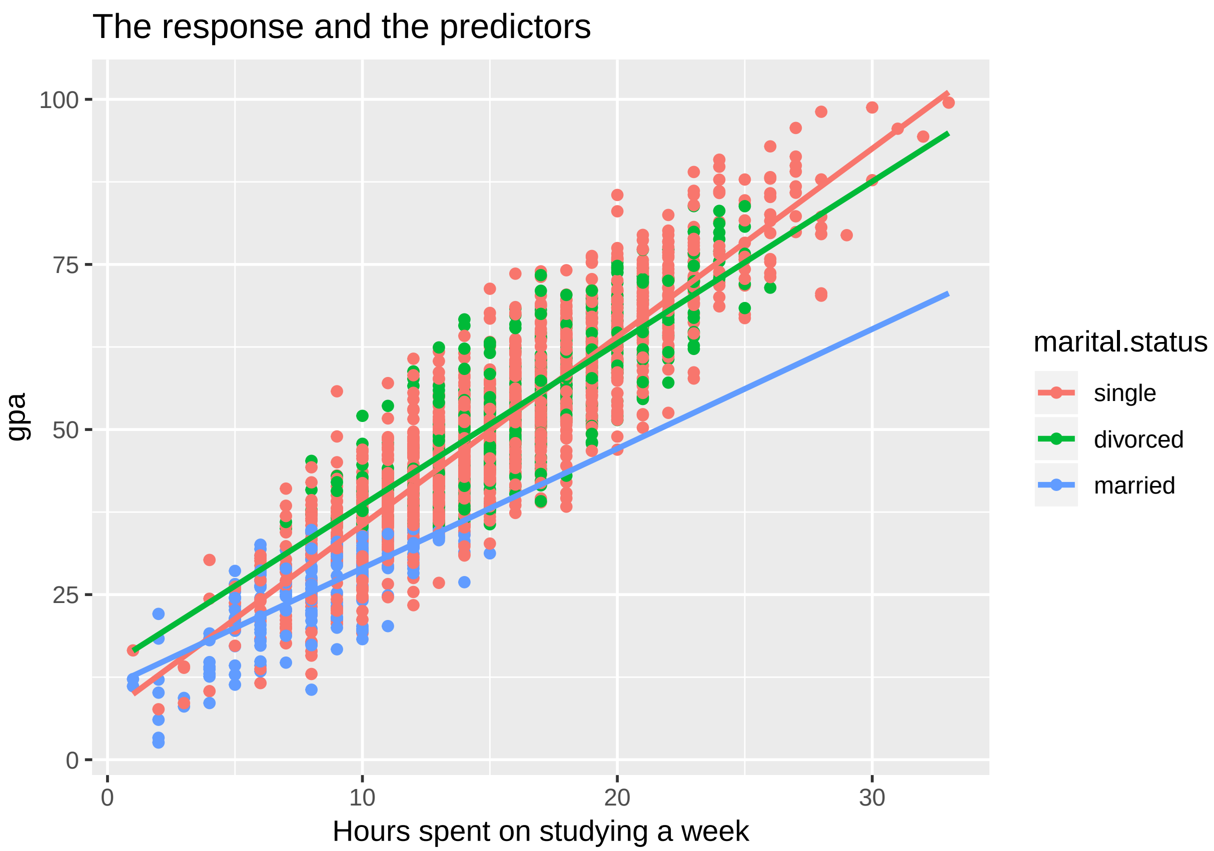
```
##
## Call:
## lm(formula = gpa ~ marital.status, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.458  -9.414  -0.313   8.226  48.402
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             51.0985     0.3708  137.82  < 2e-16 ***
## marital.statusdivorced   2.9313     0.7634    3.84 0.000127 ***
## marital.statusmarried  -25.2864     1.0561  -23.94  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13.77 on 1997 degrees of freedom
## Multiple R-squared:  0.2421, Adjusted R-squared:  0.2414
## F-statistic:   319 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```
model3<-lm(gpa~marital.status+hpw,data = Data)
library(ggiraphExtra)
ggPredict(model3,interactive=F)+ggtitle("The response and the predictors")+xlab("Hours spent on studyin
```



```
qplot(x = hpw, y = gpa, color = marital.status, data = Data) +
  stat_smooth(method = "lm", se = FALSE, fullrange = TRUE)+xlab("Hours spent on studying a week")+ggtit
```



Interquartile range of Gpa for people single and divorced more or less are similar. However, people who are married usually have lower GPA. as at least one of p-values is less than alpha, there is statistically significant relationship between gpa and marital status. As the reference category is “single”,the gpa is on average more by 2.9 for those who are divorced compared to those who are single.On the other hand married people’s gpa is on average less by 25 than gpa of single people. Now it seems that difference of gpa by single and divorced is not significant whereas married people’ Gpa quite different. Regarding plots, the one provided by method “ggPredict” do not include interaction term in order to get it you should use qplot. From the interaction qplot we can conclude that if single people increase hours spent on studying (more than 18) on average they get higher GPA.

Problem 5. 7 pt.

a. Divide the data frame into Train and Test sets (75:25). Do not forget about set.seed() function.
b. Let the threshold for correlation coefficient is 0.7. Is there multicollinearity in the data?
c. Try different models with gpa as a dependent variable. Exclude from the models one of multicollinear variables which is less correlated with GPA.
   Save only the best model (based on both R^2 and sig.t). Why do we need to look at Adjusted R^2?
d. Formulate Null and Alternative hypotheses for all model (a.k.a for F statistics).Is the H0 rejected? How do we choose the level of significance?
e. Use the stepAIC() function to obtain the best model based on AIC criteria. Use the forward selection procedure. Describe how forward selection works.
f. Calculate RMSE for the testing set for both models, which one is better.

```
set.seed(42)
index<-sample(nrow(Data),nrow(Data)*.75,replace = F)
train<-Data[index,]
test<-Data[-index,]
```

If we state threshold for correlation = 0.7,Variables Hours spent on studying a week and Hours of sleep per day are highly correlated with each other as absolute value of (-0.85)>0.75 so we should exclude one , in further model we will include only Hours spent on studying because it has stronger correlation with Gpa than hsleep has.

```
model5<-lm(gpa~.-hsleep,data= train)
summary(model5)
```

```
##
## Call:
## lm(formula = gpa ~ . - hsleep, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4985  -3.0889   0.0141   2.9570  18.3920
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             40.54491    1.60816  25.212  < 2e-16 ***
## age                      0.12102    0.05075   2.385   0.0172 *
## ehpw                    -1.39623    0.04009 -34.830  < 2e-16 ***
## hpw                      2.53578    0.04557  55.644  < 2e-16 ***
## imp.L                    1.36371    0.25721   5.302 1.32e-07 ***
## imp.Q                   -0.41036    0.26031  -1.576   0.1151
## imp.C                    0.04790    0.25852   0.185   0.8530
## imp^4                    0.16556    0.26377   0.628   0.5303
## gendermale              -2.43127    0.43775  -5.554 3.30e-08 ***
## jobUneployed             1.24444    0.54294   2.292   0.0220 *
## typepar-time            -5.11452    0.68402  -7.477 1.29e-13 ***
## marital.statusdivorced -0.51414    0.43475  -1.183   0.2372
## marital.statusmarried  -4.36553    0.50941  -8.570  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.487 on 1487 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9201
## F-statistic:  1439 on 12 and 1487 DF,  p-value: < 2.2e-16
```

The Adj R-squared increases only if including/excluding predictor improves the model. It decreases when a predictor the model by less than expected by chance. I tried different model by excluding some variables adjusted R squared decreases so I saved only this one where all predictors are significant besides correlated variables I exclude Job which was non significant and after Adj R squared remained the same. F-statistic: 1439 on 12 and 1487 degree of freedom, p-value: < 2.2e-16.The "F value" and"Prob(F)" statistics test the overall significance of the regression model. Specifically, they test the null hypothesis that all of the regression coefficients are equal to zero. This tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable. The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number. The value of Prob(F) is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero).As in our case it is less than alpha we reject it claiming that at least one slope of variables is significantly different from 0. The significance level is the probability of rejecting the null hypothesis when it is true.if we state significance level of 0.01 implies we want to be 90% percent sure that we will not reject null hypothesis when it is true.If we want to be more confident we can set alpha to 0.001

```r
summary(modAIC)
```

```
##
## Call:
## lm(formula = gpa ~ hpw + hsleep + type + marital.status + gender +
##     imp + age + job, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5370  -3.0799   0.0285   2.9748  18.3808
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -76.70656    3.00064 -25.563  < 2e-16 ***
## hpw                      3.93150    0.06337  62.045  < 2e-16 ***
## hsleep                   9.77089    0.28048  34.836  < 2e-16 ***
## typepar-time            -5.10663    0.68400  -7.466 1.40e-13 ***
## marital.statusdivorced  -0.50834    0.43472  -1.169   0.2425
## marital.statusmarried   -4.36308    0.50938  -8.565  < 2e-16 ***
## gendermale              -2.42713    0.43773  -5.545 3.48e-08 ***
## imp.L                    1.36406    0.25719   5.304 1.31e-07 ***
## imp.Q                   -0.41338    0.26029  -1.588   0.1125
## imp.C                    0.04973    0.25850   0.192   0.8475
## imp^4                    0.16628    0.26375   0.630   0.5285
## age                      0.12025    0.05074   2.370   0.0179 *
## jobUneployed             1.25262    0.54288   2.307   0.0212 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.487 on 1487 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9201
## F-statistic:  1439 on 12 and 1487 DF,  p-value: < 2.2e-16
```

Performs stepwise model selection by Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models.Direction forward it starts with intercept only model than starts adding variables that can decrease AIC score After all it will chose the model with the lowest AIC. AIC method chooses a model with correletad(hsleep and hpw) variables and excluded (ehpw).

```r
library(Metrics)
predictions1 <- predict(model5, test)
predictions2 <- predict(modAIC, test)
RMSE_MOD<-rmse(test$gpa, predictions1)
RMSE_AIC<-rmse(test$gpa, predictions2)
print(c(RMSE_MOD,RMSE_AIC))
```

```
## [1] 4.563284 4.563112
```

The AIC models provides e lower Rmse so for prediction it performs better.

Bonus. 1 pt

Do we need to prefer the OLS estimation to ML evaluation in regression models if all assumptions of Gauss-Markov theorem are satisfied for small data? Why? When we can equally use these methods?

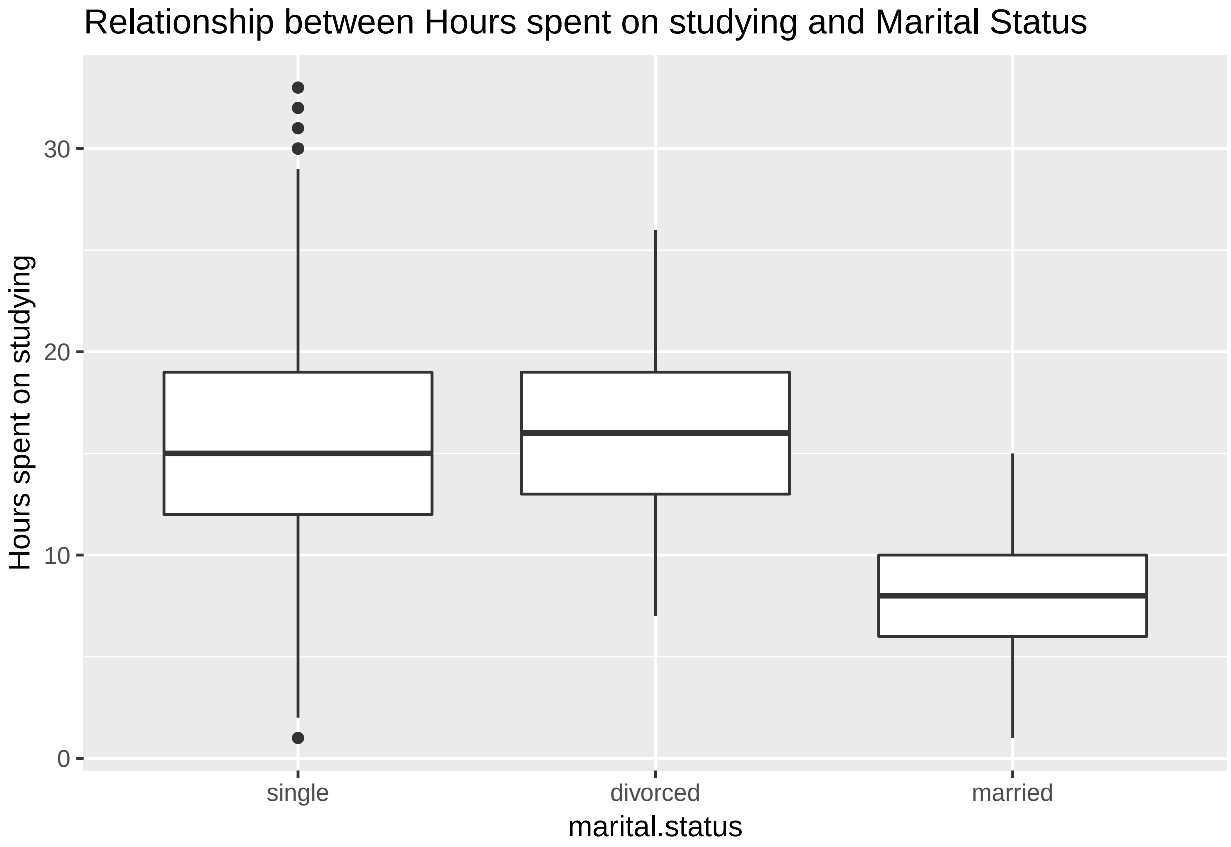
The OLS is a special case of the MLE just for normal errors and if the assumption of normality of errors is met I think both models will perform equally however if data is small and if we should use a model for predictions errors for prediction might have some pattern and we might have to use more flexible model trying maximize the likelihood The MLE is concerned about choosing the parameter which can maximize the likelihood or equivalently the log-likelihood function. And then fit the model based on the trial estimated parameter value and calculate the mean of the model. To find the iterative weighted and working dependence and based on this two and the design matrix we can estimate the best parameter value. The OLS will take the parameter value which minimize the error ( residuals) of the model. It will take into acount that the residual sum of square and derivative with respect to the parameter regression coefficient (beta) and set it to zero and then we will find the parameter value which minimize the error(residual sum of square).

One more model,please check if you have time.

```
new<-table(Data$marital.status,Data$imp)
chisq.test(new)
```

```
##
##  Pearson's Chi-squared test
##
## data:  new
## X-squared = 18.149, df = 8, p-value = 0.02014
```

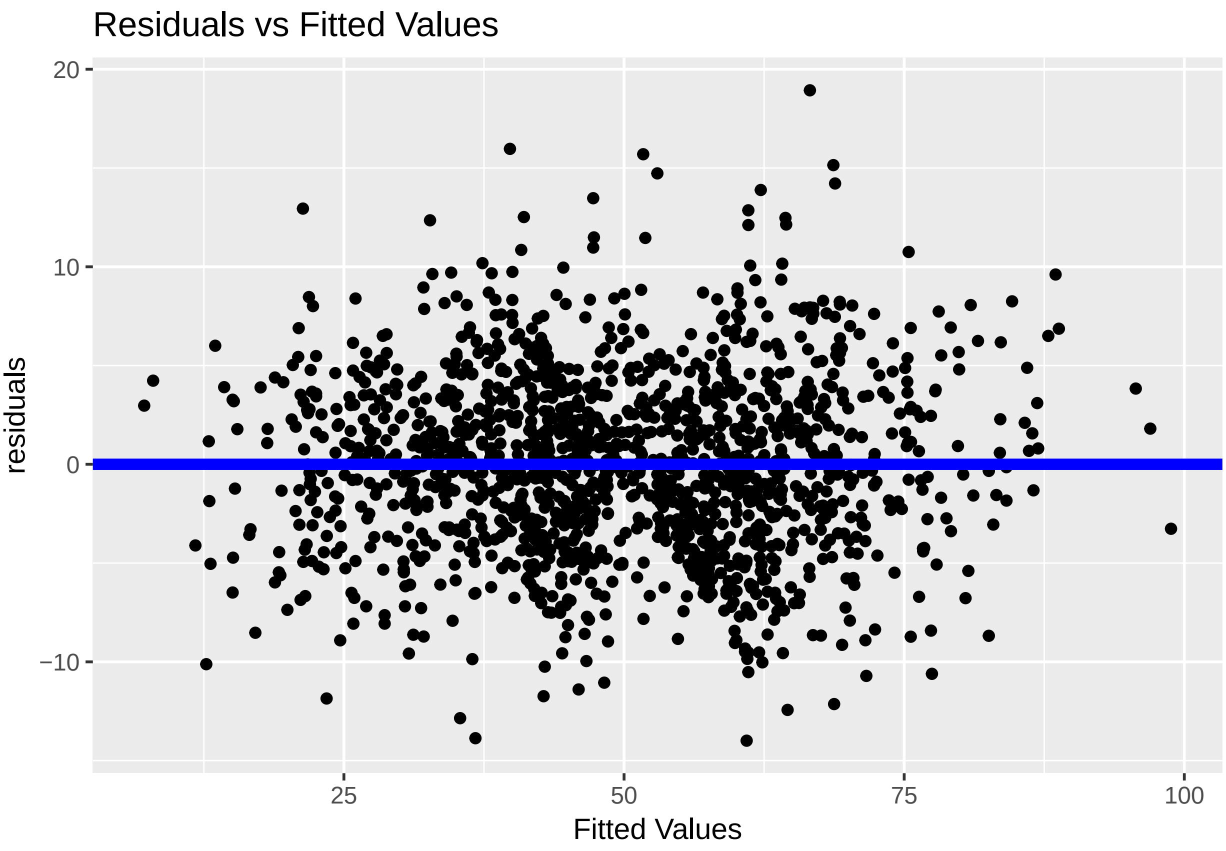
```
ggplot(Data,aes(marital.status,hpw))+geom_boxplot()+ggtitle("Relationship between Hours spent on studyi
```



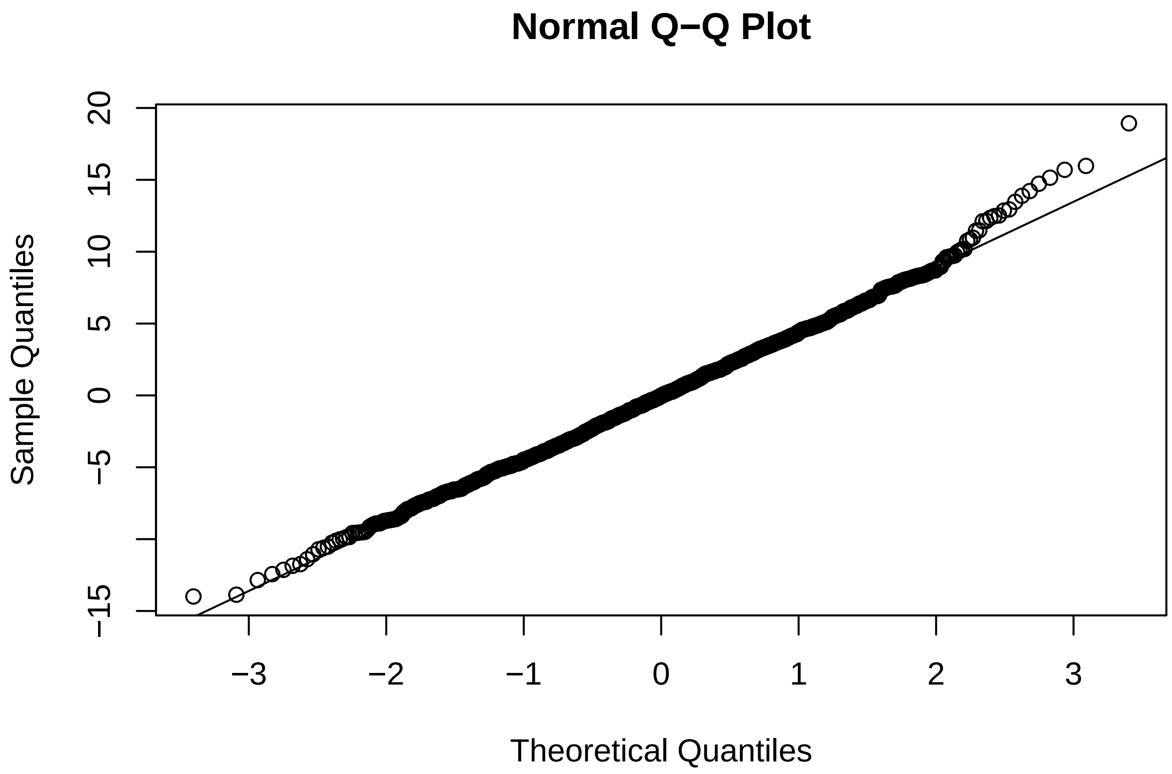
```
model6<-lm(gpa~.-hsleep-job-imp+marital.status:hpw,data= train)
summary(model6)
```

```
##
## Call:
## lm(formula = gpa ~ . - hsleep - job - imp + marital.status:hpw,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9921  -3.1116  -0.0273   2.9771  18.9344
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 41.36911    1.45838  28.366  < 2e-16 ***
## age                          0.10447    0.05107   2.046 0.040964 *
## ehpw                        -1.38477    0.04036 -34.308  < 2e-16 ***
## hpw                          2.58062    0.04853  53.172  < 2e-16 ***
## gendermale                  -2.59240    0.44583  -5.815 7.42e-09 ***
## typepar-time                -6.40601    0.41257 -15.527  < 2e-16 ***
## marital.statusdivorced       4.14290    1.32710   3.122 0.001832 **
## marital.statusmarried       -0.55853    1.31220  -0.426 0.670427
## hpw:marital.statusdivorced -0.27676    0.07585  -3.649 0.000273 ***
## hpw:marital.statusmarried  -0.39832    0.13992  -2.847 0.004477 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.506 on 1490 degrees of freedom
## Multiple R-squared:  0.9199, Adjusted R-squared:  0.9194
## F-statistic:  1901 on 9 and 1490 DF,  p-value: < 2.2e-16
```

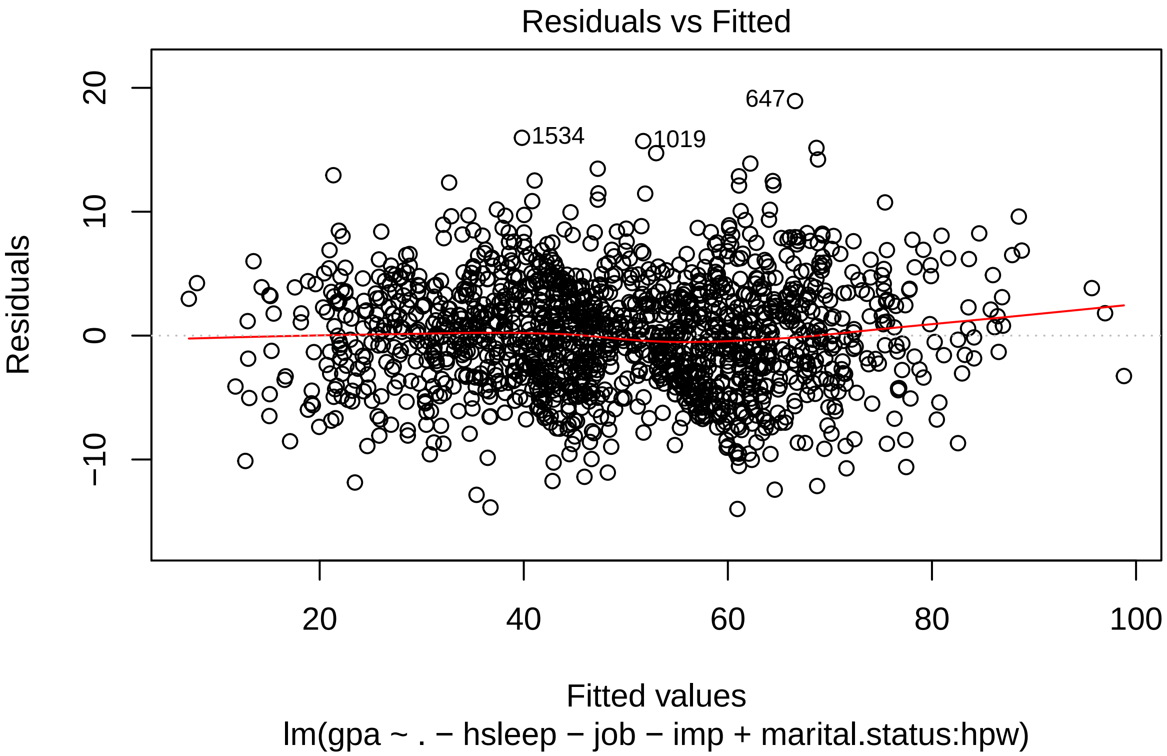
```
ggplot()+geom_point(aes(model6$fitted.values,model6$residuals))+geom_abline(intercept = 0,slope = 0,siz
```

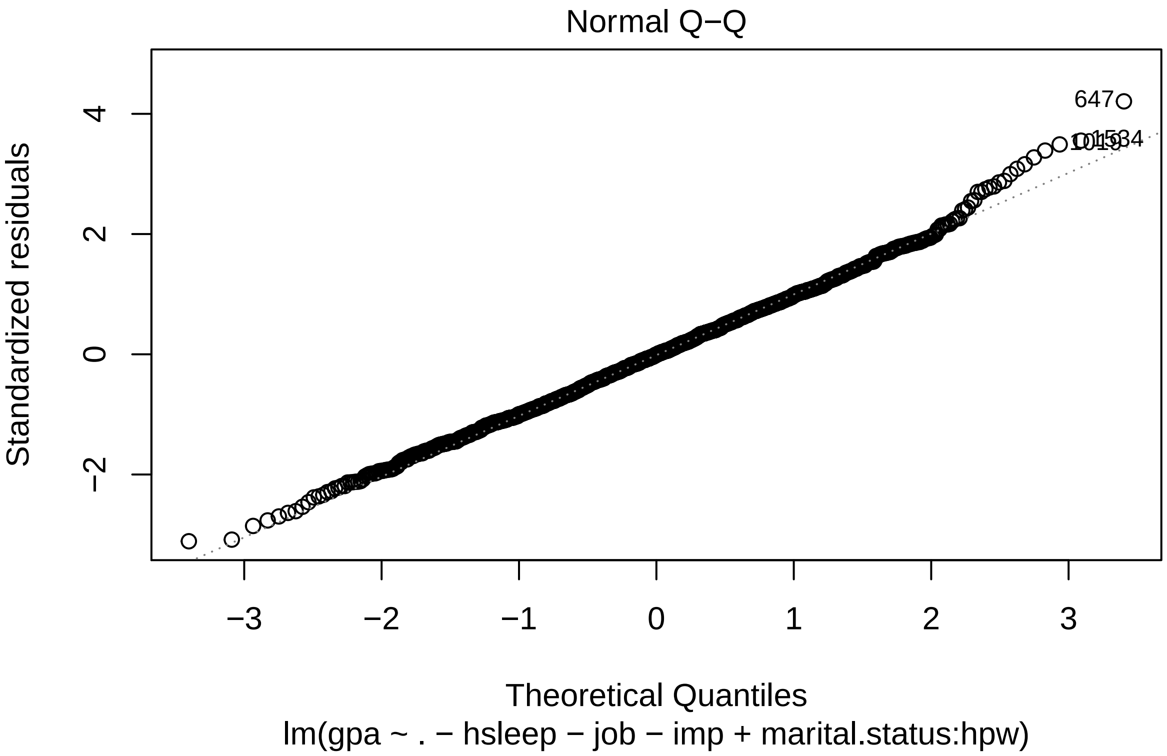


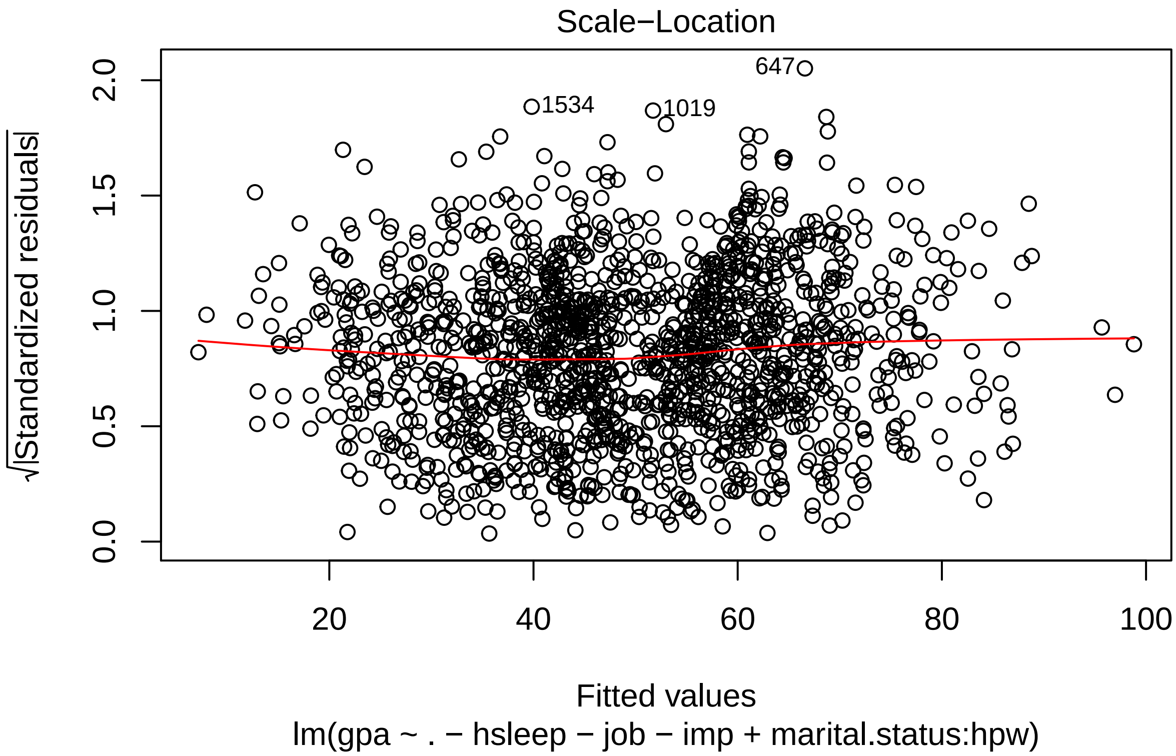
```
qqnorm(model6$residuals);qqline(model6$residuals)
```

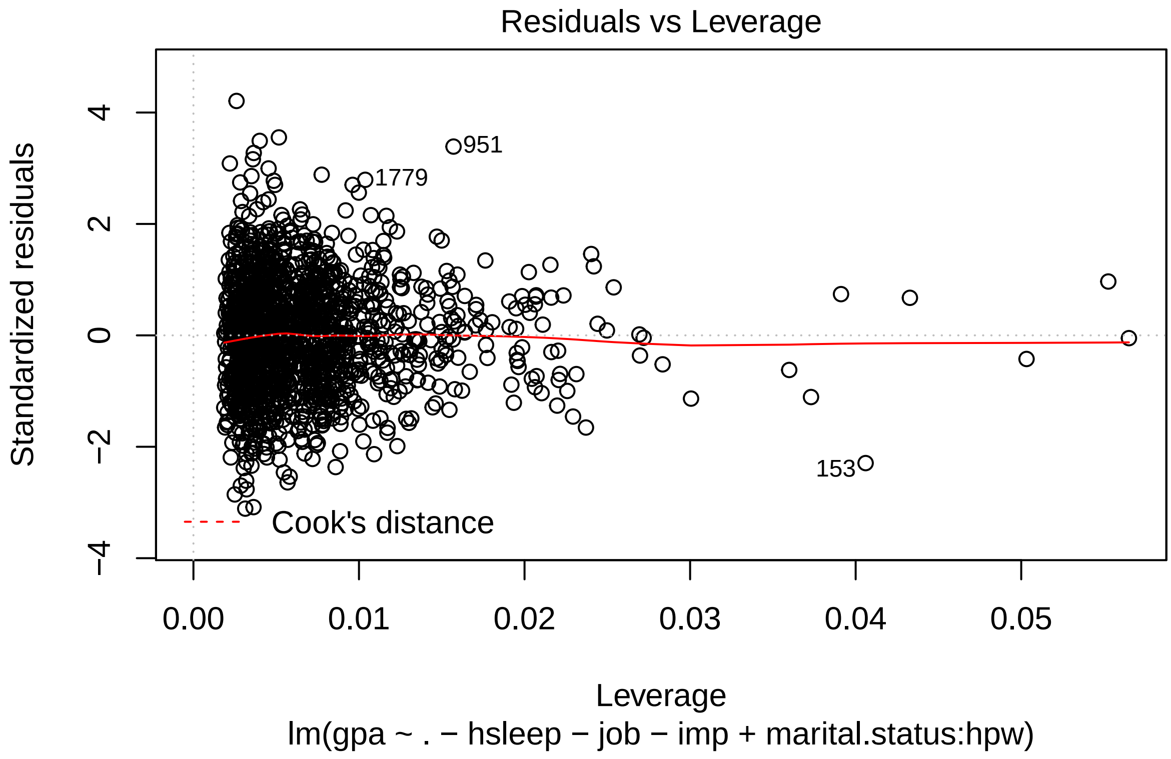


```
plot(model6)
```


Residuals vs Fitted
Residuals
Fitted values
lm(gpa ~ . − hsleep − job − imp + marital.status:hpw)
20
10
0
−10
20
40
60
80
100
647
1534
1019


Normal Q−Q
647
1015
1534
Standardized residuals
−2
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(gpa ~ . − hsleep − job − imp + marital.status:hpw)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
20
40
60
80
100
647
1534
1019
Fitted values
lm(gpa ~ . − hsleep − job − imp + marital.status:hpw)



```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
shapiro.test(model6$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model6$residuals
## W = 0.99694, p-value = 0.004905
```

```
bptest(model6)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model6
## BP = 24.373, df = 9, p-value = 0.003749
```

```
predictions3 <- predict(model6, test)
RMSE_MOD<-rmse(test$gpa, predictions3)
```

```
RMSE_MOD
```

```
## [1] 4.59818
```

As Pearson’s Chi-squared Test indicated “imp” variable and “marital.status” affect each other we need to exclude one to control independence of variables with so besides Hours of sleep per day and job the model does not include imp variable.Excluding “imp” variable from model Adjusted R squared decreased by 0.0018. Further as it was shown in box.plot there were relationship between Hours spent on studying and marital status variable so add interaction term in model and increase Adj R-squared by 0.0022.

To have best linear model

Shapiro-Wilk Normality Test H0: errors are normally distributed H:1 errors are not normally distributed.P-value is greater than alpha(0.05) so we fail to reject H0.Assumption of normality is met. Breusch-Pagan Test:H0 we have constant variance.H:1 Variance of residuals is not constant.P-value is greater than alpha so we fail to reject H0: our assumptions of homoscedasticity is met.