Module 2 Assignment

Neha Parulekar

February 1, 2016

## Assingment

* **Load the file M01\_quasi\_twitter.csv**

# Loading the M01\_quasi\_twitter.csv  
getwd()

## [1] "C:/Users/Neha/Desktop"

setwd("C:/Users/Neha/Desktop")  
QuasiTwitterData <- read.csv("M01\_quasi\_twitter.csv")  
  
# Checking the data columns and Summary   
names(QuasiTwitterData)

## [1] "screen\_name" "created\_at\_month"   
## [3] "created\_at\_day" "created\_at\_year"   
## [5] "country" "location"   
## [7] "friends\_count" "followers\_count"   
## [9] "statuses\_count" "favourites\_count"   
## [11] "favourited\_count" "dob\_day"   
## [13] "dob\_year" "dob\_month"   
## [15] "gender" "mobile\_favourites\_count"  
## [17] "mobile\_favourited\_count" "education"   
## [19] "experience" "age"   
## [21] "race" "wage"   
## [23] "retweeted\_count" "retweet\_count"   
## [25] "height"

summary(QuasiTwitterData)

## screen\_name created\_at\_month created\_at\_day created\_at\_year  
## +5400E1. : 1 Min. : 1.000 Min. : 1.00 Min. :2006   
## 000D0se7 : 1 1st Qu.: 3.000 1st Qu.: 8.00 1st Qu.:2009   
## 001apdov : 1 Median : 6.000 Median :16.00 Median :2011   
## 001RBTePh: 1 Mean : 6.069 Mean :15.78 Mean :2011   
## 003B0K2 : 1 3rd Qu.: 9.000 3rd Qu.:23.00 3rd Qu.:2013   
## 007unfasa: 1 Max. :12.000 Max. :31.00 Max. :2015   
## (Other) :21910   
## country location friends\_count   
## USA :14905 Mexico : 122 Min. : -84   
## Canada : 943 Boston : 108 1st Qu.: 123   
## India : 890 Montreal : 107 Median : 324   
## Earth : 516 Nevada : 80 Mean : 1058   
## England : 467 Bangalore : 79 3rd Qu.: 849   
## Australia: 291 Indianapolis Indiana: 76 Max. :660549   
## (Other) : 3904 (Other) :21344   
## followers\_count statuses\_count favourites\_count   
## Min. : 0 Min. : 1 Min. : 0   
## 1st Qu.: 105 1st Qu.: 558 1st Qu.: 16   
## Median : 336 Median : 2341 Median : 164   
## Mean : 5859 Mean : 12486 Mean : 2217   
## 3rd Qu.: 1075 3rd Qu.: 9348 3rd Qu.: 950   
## Max. :22187643 Max. :1136198 Max. :1140139   
##   
## favourited\_count dob\_day dob\_year dob\_month   
## Min. : 0.00 Min. : 1.00 Min. :1900 Min. : 1.000   
## 1st Qu.: 2.00 1st Qu.: 5.00 1st Qu.:1965 1st Qu.: 3.000   
## Median : 9.00 Median :13.00 Median :1982 Median : 6.000   
## Mean : 92.24 Mean :13.49 Mean :1976 Mean : 6.398   
## 3rd Qu.: 36.00 3rd Qu.:21.00 3rd Qu.:1990 3rd Qu.: 9.000   
## Max. :105005.00 Max. :35.00 Max. :2000 Max. :1992.000   
##   
## gender mobile\_favourites\_count mobile\_favourited\_count  
## female: 7319 Min. : 0.0 Min. : 0   
## male :14569 1st Qu.: 0.0 1st Qu.: 0   
## NA's : 28 Median : 0.0 Median : 0   
## Mean : 152.9 Mean : 649   
## 3rd Qu.: 0.0 3rd Qu.: 0   
## Max. :377123.0 Max. :5032191   
##   
## education experience age race   
## Min. : 3.0 Min. :-32.00 Min. :-6.00 white :18032   
## 1st Qu.:11.0 1st Qu.: 0.00 1st Qu.:28.00 latino : 1115   
## Median :13.0 Median : 7.00 Median :36.00 asian : 960   
## Mean :12.5 Mean : 10.88 Mean :35.54 persian : 376   
## 3rd Qu.:14.0 3rd Qu.: 20.00 3rd Qu.:44.00 hispanic : 353   
## Max. :24.0 Max. : 74.00 Max. :91.00 pacific islander: 276   
## (Other) : 804   
## wage retweeted\_count retweet\_count height   
## Min. : 5.00 Min. : 0.0000 Min. : 0.00 Min. : 1.0   
## 1st Qu.: 13.52 1st Qu.: 0.0000 1st Qu.: 0.00 1st Qu.:165.0   
## Median : 20.36 Median : 1.0000 Median : 3.00 Median :172.0   
## Mean : 22.97 Mean : 0.9715 Mean : 52.73 Mean :171.5   
## 3rd Qu.: 28.40 3rd Qu.: 1.0000 3rd Qu.: 19.00 3rd Qu.:178.0   
## Max. :104.97 Max. :705.0000 Max. :5506.00 Max. :203.0   
##

## Additional packages needed

To run the code in Module2L2.Rmd you may need the following packages.

* If necessary install ggplot2 package.

install.packages("ggplot2");install.packages("reshape2")

require(ggplot2)

## Loading required package: ggplot2

require(reshape2)

## Loading required package: reshape2

* **Generate a linear model for the following:**
  + A relation between followers\_count & gender
  + A relation between dob\_year & statuses\_count
  + A significant linear linear model of your choosing.
  + A multivariate relation between wage & height, race, age, education & experience
  + A significant logistic linear model of your choosing.
* **Answer the following questions:**
  + Is the relationship significant?
  + Are any model assumptions violated?
  + Is there any multi-colinearity in multivariate models?
  + In in multivariate models are predictor variables independent of all the other predictor variables?
  + In in multivariate models rank the most significant predictor variables and exclude insignificant one from the model.
  + Does the model make sense?

## Answers

* **A relation between followers\_count & gender**

# getting the relationship  
m\_followers\_count\_gender <- lm(followers\_count ~ gender, data = QuasiTwitterData )  
m\_followers\_count\_gender

##   
## Call:  
## lm(formula = followers\_count ~ gender, data = QuasiTwitterData)  
##   
## Coefficients:  
## (Intercept) gendermale   
## 8580 -4077

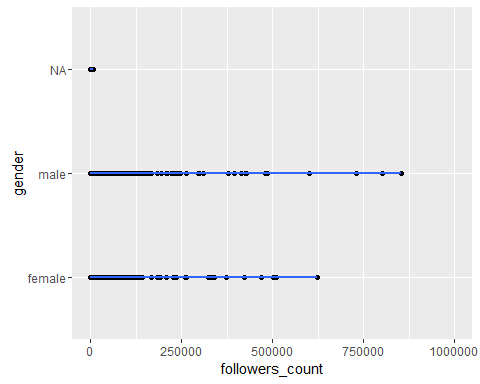
# getting the summary  
summary(m\_followers\_count\_gender)

##   
## Call:  
## lm(formula = followers\_count ~ gender, data = QuasiTwitterData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8580 -7443 -4373 -3823 22179063   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8580 1997 4.297 1.74e-05 \*\*\*  
## gendermale -4077 2447 -1.666 0.0958 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 170800 on 21886 degrees of freedom  
## (28 observations deleted due to missingness)  
## Multiple R-squared: 0.0001268, Adjusted R-squared: 8.109e-05   
## F-statistic: 2.775 on 1 and 21886 DF, p-value: 0.09576

# plotting the data   
qplot(followers\_count, gender, data = QuasiTwitterData) + geom\_smooth(method = lm) + ylab("gender") + xlab("followers\_count") + xlim(-100, 1000000)

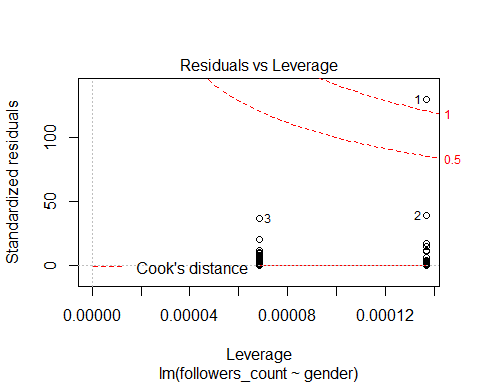
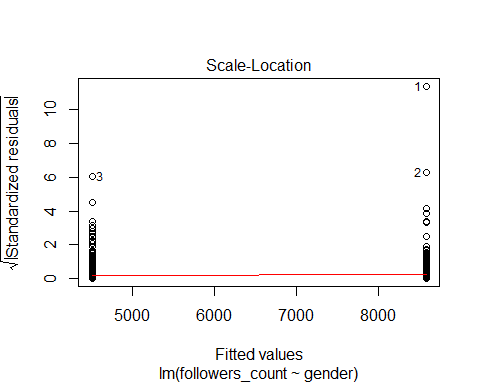
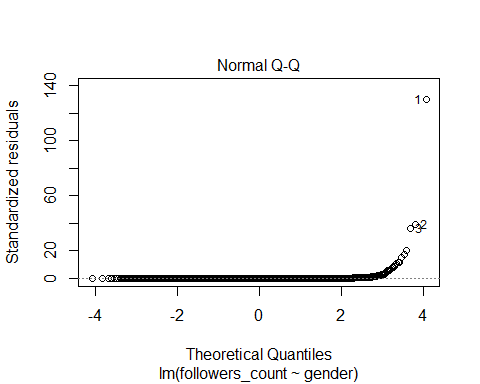
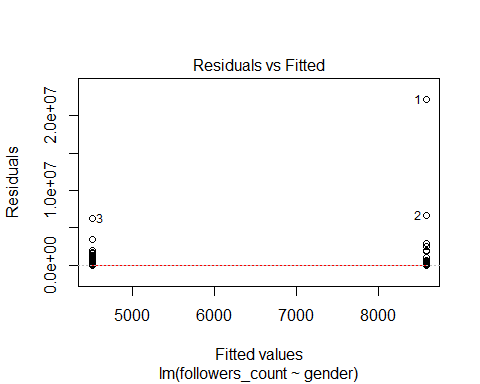
## Warning: Removed 20 rows containing non-finite values (stat\_smooth).

## Warning: Removed 20 rows containing missing values (geom\_point).

 **Is the relationship significant?** We cannot say that the relationship is significant solely based on the gender and follower counts as there are other factor like friend count, age etc which also play major part in knowing proper follower count.

**Are any model assumptions violated?** Passing the model to the plot function we see that model assumptions are violated

plot(m\_followers\_count\_gender)

 We can see the outliers in residual leverage plot. There is no constant variance, Homoscedasticity.Normality assunption is violated.

**Does the model make sense?** No, the model does not make sense as there are other variables which have to be taken into account to have any significance.

* **A relation between dob\_year & statuses\_count**

# getting the relationship  
m\_dob\_year\_statuses\_count <- lm(dob\_year ~ statuses\_count, data = QuasiTwitterData)  
m\_dob\_year\_statuses\_count

##   
## Call:  
## lm(formula = dob\_year ~ statuses\_count, data = QuasiTwitterData)  
##   
## Coefficients:  
## (Intercept) statuses\_count   
## 1.976e+03 1.133e-05

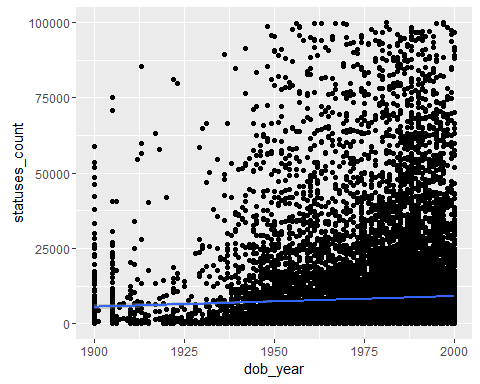
# getting the summary   
summary(m\_dob\_year\_statuses\_count)

##   
## Call:  
## lm(formula = dob\_year ~ statuses\_count, data = QuasiTwitterData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -77.969 -11.105 5.888 13.983 23.997   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.976e+03 1.361e-01 14516.049 < 2e-16 \*\*\*  
## statuses\_count 1.133e-05 3.557e-06 3.185 0.00145 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.05 on 21914 degrees of freedom  
## Multiple R-squared: 0.0004626, Adjusted R-squared: 0.000417   
## F-statistic: 10.14 on 1 and 21914 DF, p-value: 0.001451

# plotting the data  
qplot(dob\_year, statuses\_count, data = QuasiTwitterData) + geom\_smooth(method = lm) + xlab("dob\_year") + ylab("statuses\_count") + ylim(-100, 100000)

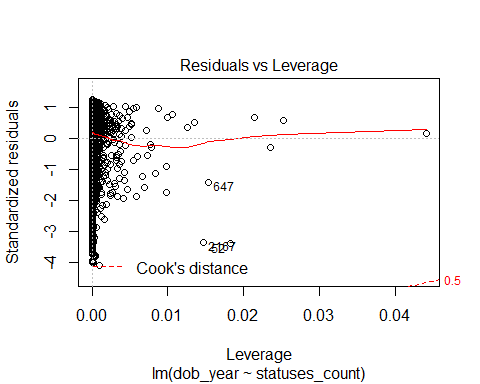
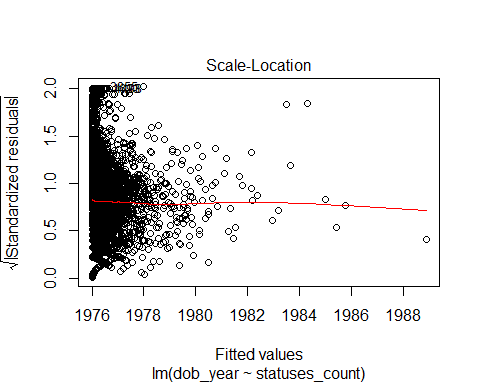
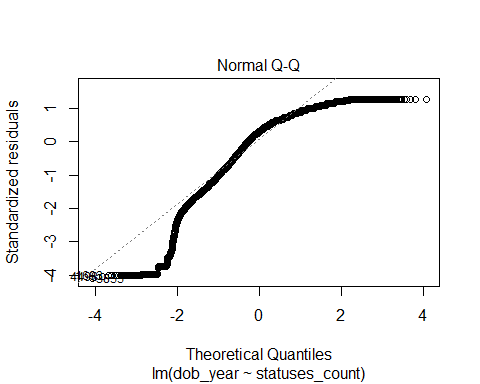
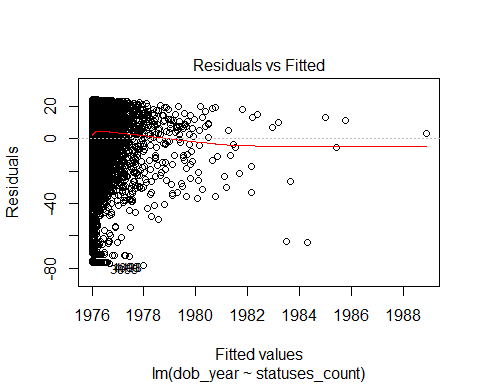
## Warning: Removed 501 rows containing non-finite values (stat\_smooth).

## Warning: Removed 501 rows containing missing values (geom\_point).

 **Is the relationship significant?** The relationship is not significant as the statuses count should not have any relation with the year a person was born. There are many factors which have to be taken into consideration.

**Are any model assumptions violated?** Passing the model to the plot function we see that model assumptions are violated

plot(m\_dob\_year\_statuses\_count)

 we can state that there is no constanat variance. we can see outliers in residual leverage plot.

**Does the model make sense?** The model does not make sense as there are many factors which have not been taken into account.

* **A significant linear model of your choosing.** Plotting a linear model between experience and wage

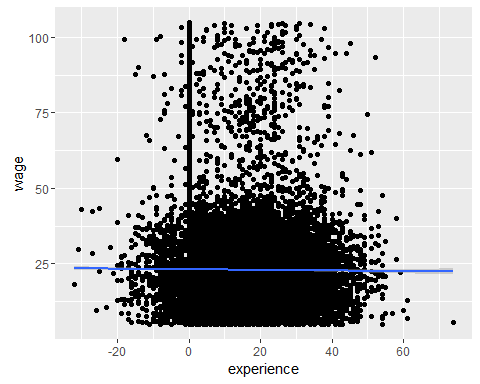
# getting the relationship   
m\_wage\_experience <- lm(wage ~ experience, data = QuasiTwitterData)   
m\_wage\_experience

##   
## Call:  
## lm(formula = wage ~ experience, data = QuasiTwitterData)  
##   
## Coefficients:  
## (Intercept) experience   
## 23.07994 -0.00983

# getting the summary   
summary(QuasiTwitterData)

## screen\_name created\_at\_month created\_at\_day created\_at\_year  
## +5400E1. : 1 Min. : 1.000 Min. : 1.00 Min. :2006   
## 000D0se7 : 1 1st Qu.: 3.000 1st Qu.: 8.00 1st Qu.:2009   
## 001apdov : 1 Median : 6.000 Median :16.00 Median :2011   
## 001RBTePh: 1 Mean : 6.069 Mean :15.78 Mean :2011   
## 003B0K2 : 1 3rd Qu.: 9.000 3rd Qu.:23.00 3rd Qu.:2013   
## 007unfasa: 1 Max. :12.000 Max. :31.00 Max. :2015   
## (Other) :21910   
## country location friends\_count   
## USA :14905 Mexico : 122 Min. : -84   
## Canada : 943 Boston : 108 1st Qu.: 123   
## India : 890 Montreal : 107 Median : 324   
## Earth : 516 Nevada : 80 Mean : 1058   
## England : 467 Bangalore : 79 3rd Qu.: 849   
## Australia: 291 Indianapolis Indiana: 76 Max. :660549   
## (Other) : 3904 (Other) :21344   
## followers\_count statuses\_count favourites\_count   
## Min. : 0 Min. : 1 Min. : 0   
## 1st Qu.: 105 1st Qu.: 558 1st Qu.: 16   
## Median : 336 Median : 2341 Median : 164   
## Mean : 5859 Mean : 12486 Mean : 2217   
## 3rd Qu.: 1075 3rd Qu.: 9348 3rd Qu.: 950   
## Max. :22187643 Max. :1136198 Max. :1140139   
##   
## favourited\_count dob\_day dob\_year dob\_month   
## Min. : 0.00 Min. : 1.00 Min. :1900 Min. : 1.000   
## 1st Qu.: 2.00 1st Qu.: 5.00 1st Qu.:1965 1st Qu.: 3.000   
## Median : 9.00 Median :13.00 Median :1982 Median : 6.000   
## Mean : 92.24 Mean :13.49 Mean :1976 Mean : 6.398   
## 3rd Qu.: 36.00 3rd Qu.:21.00 3rd Qu.:1990 3rd Qu.: 9.000   
## Max. :105005.00 Max. :35.00 Max. :2000 Max. :1992.000   
##   
## gender mobile\_favourites\_count mobile\_favourited\_count  
## female: 7319 Min. : 0.0 Min. : 0   
## male :14569 1st Qu.: 0.0 1st Qu.: 0   
## NA's : 28 Median : 0.0 Median : 0   
## Mean : 152.9 Mean : 649   
## 3rd Qu.: 0.0 3rd Qu.: 0   
## Max. :377123.0 Max. :5032191   
##   
## education experience age race   
## Min. : 3.0 Min. :-32.00 Min. :-6.00 white :18032   
## 1st Qu.:11.0 1st Qu.: 0.00 1st Qu.:28.00 latino : 1115   
## Median :13.0 Median : 7.00 Median :36.00 asian : 960   
## Mean :12.5 Mean : 10.88 Mean :35.54 persian : 376   
## 3rd Qu.:14.0 3rd Qu.: 20.00 3rd Qu.:44.00 hispanic : 353   
## Max. :24.0 Max. : 74.00 Max. :91.00 pacific islander: 276   
## (Other) : 804   
## wage retweeted\_count retweet\_count height   
## Min. : 5.00 Min. : 0.0000 Min. : 0.00 Min. : 1.0   
## 1st Qu.: 13.52 1st Qu.: 0.0000 1st Qu.: 0.00 1st Qu.:165.0   
## Median : 20.36 Median : 1.0000 Median : 3.00 Median :172.0   
## Mean : 22.97 Mean : 0.9715 Mean : 52.73 Mean :171.5   
## 3rd Qu.: 28.40 3rd Qu.: 1.0000 3rd Qu.: 19.00 3rd Qu.:178.0   
## Max. :104.97 Max. :705.0000 Max. :5506.00 Max. :203.0   
##

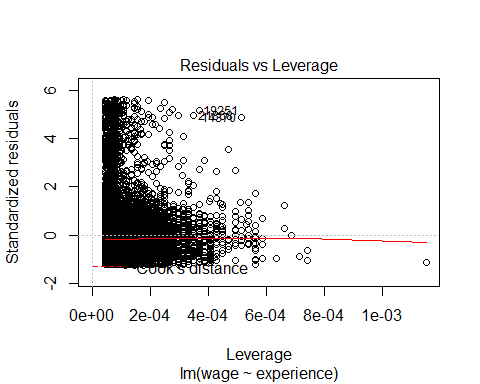
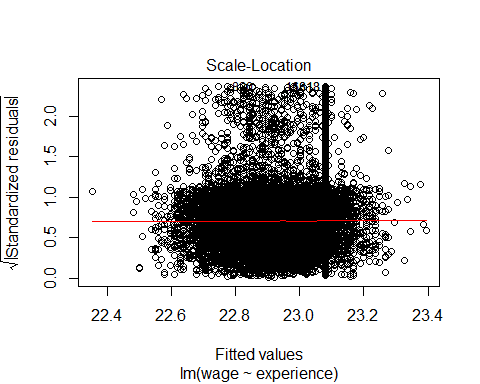
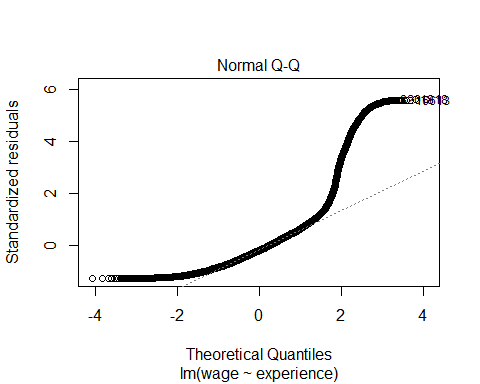
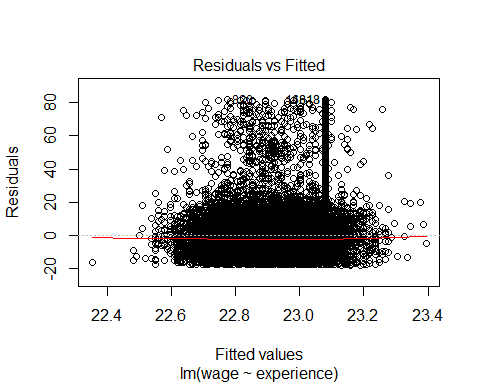
# plotting the data  
qplot(experience, wage, data = QuasiTwitterData) + geom\_smooth(method = lm) + xlab("experience") + ylab("wage")



**Is the relationship significant?** We would generally assume the wages to increase with increase in experience.In the plot we can see a negative slope which means the wages decreases as the experience is increasing. Also we can see that the experience has negative years which is not possible. Thus we cannot reject the null hypothesis, and can conclude that the relationship is **not** significant

**Are any model assumptions violated?** Passing the model to the plot function we see that there are model assumptions violated.

plot(m\_wage\_experience)

 Like: Normality assumption is violated. outliers can be observed in residual leverage plot. Also, there is no constant variance.

**Does the model make sense?** The model does not make sense as there are cases where young people who lack experience could be paid more as the data like education, country ect would also effect the wage.

* **A multivariate relation between wage & height, race, age, education & experience.**

m\_wage\_height\_race\_age\_education\_experience <- lm(wage ~ height + race + age + education + experience, data = QuasiTwitterData)  
m\_wage\_height\_race\_age\_education\_experience

##   
## Call:  
## lm(formula = wage ~ height + race + age + education + experience,   
## data = QuasiTwitterData)  
##   
## Coefficients:  
## (Intercept) height raceasian   
## -12.003381 0.193717 0.911786   
## racehispanic raceindian racelatino   
## 1.730412 1.400118 1.057495   
## racemixed racenative american racepacific islander   
## 0.092136 0.513804 0.392584   
## racepersian racewhite age   
## 1.325729 0.865941 0.004442   
## education experience   
## 0.060723 -0.004366

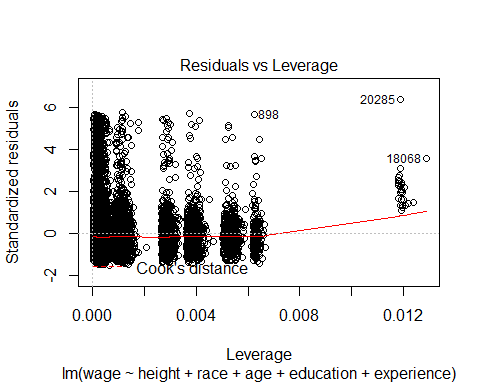
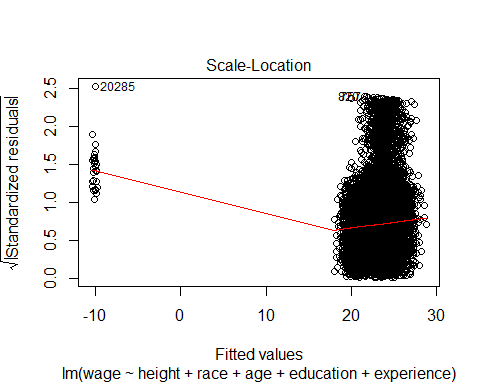
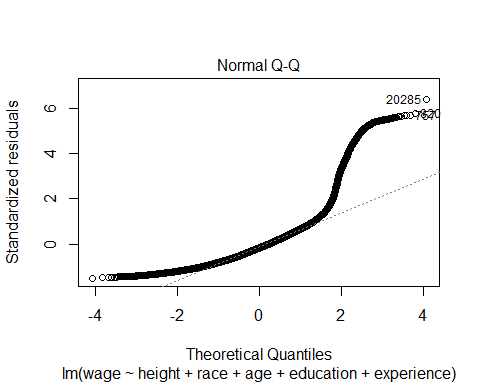
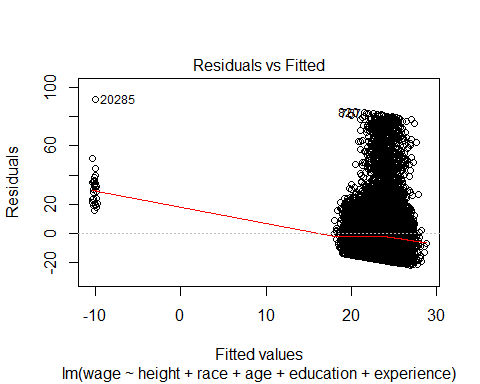
anova(m\_wage\_height\_race\_age\_education\_experience)

## Analysis of Variance Table  
##   
## Response: wage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## height 1 93912 93912 449.1906 <2e-16 \*\*\*  
## race 9 777 86 0.4127 0.9292   
## age 1 69 69 0.3317 0.5647   
## education 1 565 565 2.7033 0.1002   
## experience 1 68 68 0.3267 0.5676   
## Residuals 21902 4579052 209   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Is the relationship significant?** When compairing the wages and other variables, we can correlate the wages with other variables and rank them. Thus the relationship is significant. Also the experence cannot be negative as seen above.

**Are any model assumptions violated?**

# getting the residual plots for the model  
plot(m\_wage\_height\_race\_age\_education\_experience)

 From the plots above we can say that \* Variables are not independent statistically \* There is no constant variance \* Normality assumption is violated So we can conclude that model assumptions are violated.

**Is there any multi-colinearity in multivariate models?**

#head(lm.influence(m\_wage\_height\_race\_age\_education\_experience))

Age, experience and edcation seem to be correlated with each other.

**In in multivariate models are predictor variables independent of all the other predictor variables?** No, in case of age, experience and education, these variables are not statistically independent.

**In in multivariate models rank the most significant predictor variables and exclude insignificant one from the model.**

SubsetVariableData <- subset(QuasiTwitterData, gender == "male", select = c(wage, height, race, age, education, experience))  
beg <- lm(wage ~ height, data = SubsetVariableData)  
end <- lm(wage ~ ., data = SubsetVariableData)  
empty <- empty<-lm(wage ~ 1, data = SubsetVariableData)  
bounds <- list(upper=end, lower=empty)  
stepwise\_reg <- step(beg, bounds, direction="forward")

## Start: AIC=81401.93  
## wage ~ height  
##   
## Df Sum of Sq RSS AIC  
## <none> 3889204 81402  
## + education 1 533.08 3888671 81402  
## + age 1 468.04 3888736 81402  
## + experience 1 24.03 3889180 81404  
## + race 9 1268.96 3887935 81415

According to the AIC, the most significant variables are experience followed by age and education

**Does the model make sense?** The model makes perfect sense as we can see that wage is altered based on the age, education, experience as we would expect.

* **A significant logistic linear model of your choosing.**

Logistic regression measures the relationship between the categorical response variable and one or more predictor variables by estimating probabilities. Often this is used when the variable is binary Can height predict gender?

# logistic linear model   
m\_logistic\_model\_gender <- lm(as.numeric(gender) ~ height, data = QuasiTwitterData)  
m\_logistic\_model\_gender

##   
## Call:  
## lm(formula = as.numeric(gender) ~ height, data = QuasiTwitterData)  
##   
## Coefficients:  
## (Intercept) height   
## -5.33007 0.04073

# getting the summary   
summary(m\_logistic\_model\_gender)

##   
## Call:  
## lm(formula = as.numeric(gender) ~ height, data = QuasiTwitterData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.96104 -0.22785 -0.00177 0.24263 1.13874   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.3300696 0.0411321 -129.6 <2e-16 \*\*\*  
## height 0.0407324 0.0002392 170.3 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3094 on 21886 degrees of freedom  
## (28 observations deleted due to missingness)  
## Multiple R-squared: 0.5699, Adjusted R-squared: 0.5699   
## F-statistic: 2.9e+04 on 1 and 21886 DF, p-value: < 2.2e-16

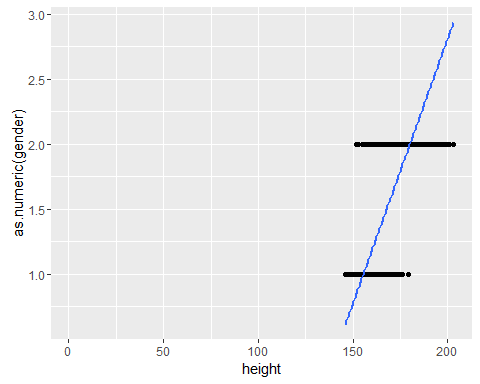
# Analysis of Variance Table  
anova(m\_logistic\_model\_gender)

## Analysis of Variance Table  
##   
## Response: as.numeric(gender)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## height 1 2776.4 2776.4 29002 < 2.2e-16 \*\*\*  
## Residuals 21886 2095.2 0.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Plotting   
qplot(height, as.numeric(gender), data = QuasiTwitterData) + stat\_smooth(method = "lm", formula = y ~ x, se = FALSE)

## Warning: Removed 28 rows containing non-finite values (stat\_smooth).

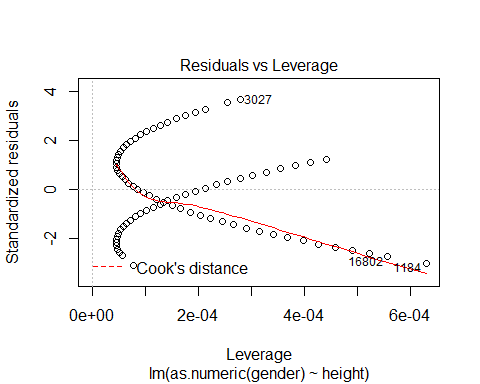
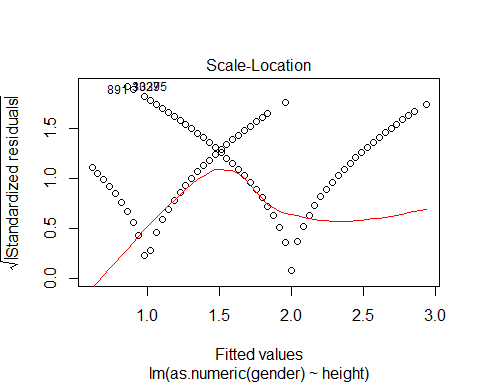
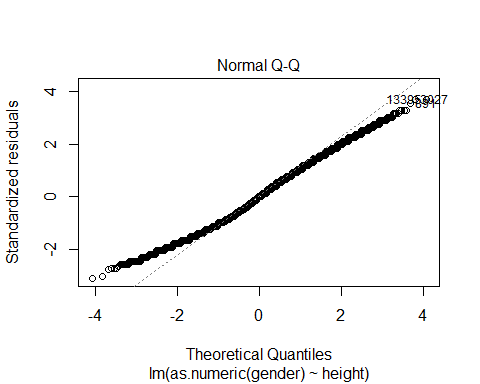
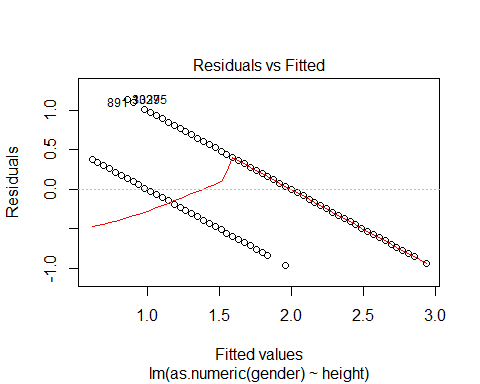
## Warning: Removed 28 rows containing missing values (geom\_point).



**Is the relationship significant?** We generally assume that the males would be taller than the females. Thus the relationship would be significant.

**Are any model assumptions violated?** Passing the model to the plot function we see that there are **NO** model assumptions violated.

plot(m\_logistic\_model\_gender)



**Does the model make sense?** As we can interpret, the females height range is less than that of males, which was as assumed and the output is binary, we can say that the model makes sense.