Module2\_assignment

Xueyi Fan

May 21, 2016

This homework assignment focuses on Linear Models, Model Inference and Interpretation. You will provide a written analysis based on the following information:

* First, load the file M01\_quasi\_twitter.csv (it is online at <http://nikbearbrown.com/YouTube/MachineLearning/M01/M01_quasi_twitter.csv>'). Next, generate a linear model for the following:
  + A relation between followers\_count & gender
  + A relation between dob\_year & statuses\_count
  + Two significant linear models of your choosing (Note that you may need to try a few models before finding two significant linear models)
  + A multivariate relation between wage & height, race, age, education, and experience
  + A significant logistic linear model of your choosing
* Finally, answer the following questions:
  1. Is the relationship significant?
  2. Are any model assumptions violated?
  3. Is there any multi-colinearity in multivariate models?
  4. In in multivariate models are predictor variables independent of all the other predictor variables?
  5. In in multivariate models rank the most significant predictor variables and exclude insignificant one from the model.
  6. Does the model make sense? Why or why not?

## To load the file M01\_quasi\_twitter.csv

require("ggplot2")

## Loading required package: ggplot2

require("reshape2")

## Loading required package: reshape2

data\_url <-'http://nikbearbrown.com/YouTube/MachineLearning/M01/M01\_quasi\_twitter.csv'  
twitter\_data<- read.csv( url(data\_url) )  
str(twitter\_data)

## 'data.frame': 21916 obs. of 25 variables:  
## $ screen\_name : Factor w/ 21916 levels "+5400E1.","000D0se7",..: 4341 15303 21127 13570 14085 3607 14942 8653 15547 19146 ...  
## $ created\_at\_month : int 2 11 4 3 4 2 7 5 1 1 ...  
## $ created\_at\_day : int 9 21 1 24 23 9 15 23 23 13 ...  
## $ created\_at\_year : int 2007 2009 2007 2007 2009 2009 2006 2008 2009 2009 ...  
## $ country : Factor w/ 44 levels " Germany","Argentina",..: 44 19 19 44 44 12 44 5 44 44 ...  
## $ location : Factor w/ 378 levels "Akron Ohio","Alabama",..: 188 202 25 233 211 79 365 41 242 83 ...  
## $ friends\_count : int 1087 5210 1015 338 641 917 1574 16300 8316 640 ...  
## $ followers\_count : int 22187643 6692814 6257020 3433218 2929559 2540842 1960373 1934803 1855827 1697620 ...  
## $ statuses\_count : int 60246 93910 118465 78082 93892 59397 41023 62178 56057 82912 ...  
## $ favourites\_count : int 1122 3825 1143 0 226 2122 20160 15 540 3 ...  
## $ favourited\_count : int 105005 40487 87968 25943 32589 19760 13558 25084 8732 24515 ...  
## $ dob\_day : int 29 24 4 22 9 1 2 6 15 26 ...  
## $ dob\_year : int 1999 1991 1997 1998 1963 1995 1999 1986 1991 1986 ...  
## $ dob\_month : int 4 10 3 8 11 1 11 10 2 9 ...  
## $ gender : Factor w/ 2 levels "female","male": 1 1 2 2 1 1 1 2 1 2 ...  
## $ mobile\_favourites\_count: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ mobile\_favourited\_count: int 0 5032191 0 0 0 0 0 1934803 0 0 ...  
## $ education : int 8 15 9 9 13 15 14 10 11 12 ...  
## $ experience : int 0 0 0 44 24 21 31 0 27 20 ...  
## $ age : int 29 0 32 40 45 14 27 31 34 40 ...  
## $ race : Factor w/ 10 levels "arab","asian",..: 10 10 10 10 10 10 10 10 2 1 ...  
## $ wage : num 16.3 17.9 15.7 7 17.9 ...  
## $ retweeted\_count : int 1 1 2 0 1 2 1 2 0 0 ...  
## $ retweet\_count : int 30 6 65 8 7 64 13 14 15 10 ...  
## $ height : int 156 162 168 180 162 158 160 178 156 173 ...

names(twitter\_data)

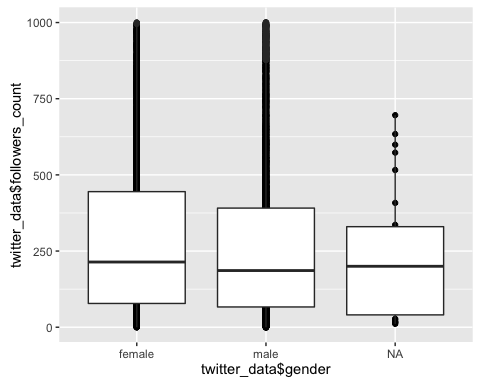
## [1] "screen\_name" "created\_at\_month"   
## [3] "created\_at\_day" "created\_at\_year"   
## [5] "country" "location"   
## [7] "friends\_count" "followers\_count"   
## [9] "statuses\_count" "favourites\_count"   
## [11] "favourited\_count" "dob\_day"   
## [13] "dob\_year" "dob\_month"   
##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

## A relation between followers\_count & gender

qplot( twitter\_data$gender, twitter\_data$followers\_count) + geom\_boxplot() + ylim(0,1000)

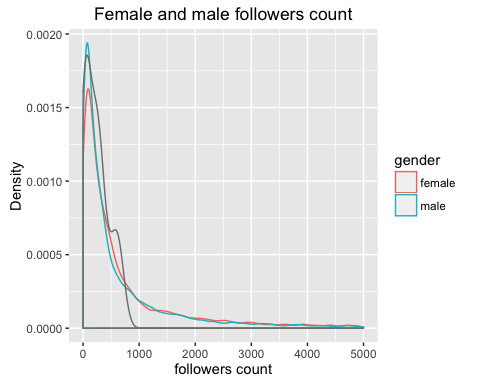
## Warning: Removed 5753 rows containing non-finite values (stat\_boxplot).

## Warning: Removed 5753 rows containing missing values (geom\_point).



ggplot(twitter\_data, aes(followers\_count, colour= gender)) + geom\_density()+xlim (0,5\*10^3) + labs(title = "Female and male followers count", y = "Density", x= "followers count")

## Warning: Removed 1870 rows containing non-finite values (stat\_density).



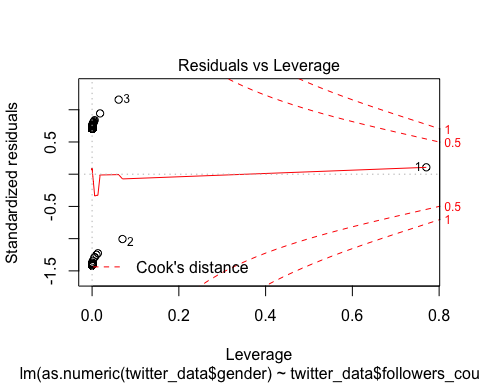
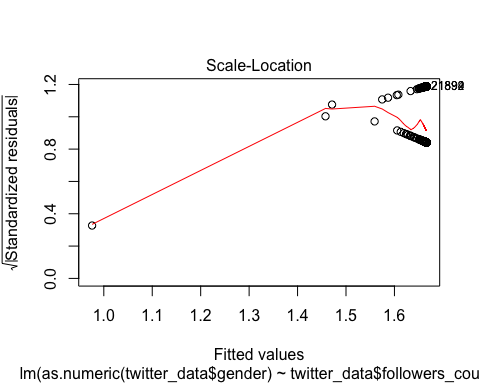
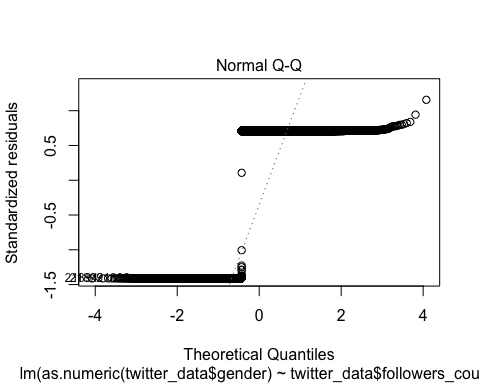
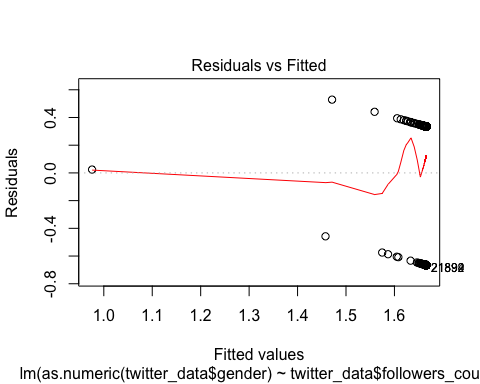
m\_follower\_gender<- lm(as.numeric(twitter\_data$gender)~twitter\_data$followers\_count)  
  
summary(m\_follower\_gender)

##   
## Call:  
## lm(formula = as.numeric(twitter\_data$gender) ~ twitter\_data$followers\_count)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.6658 -0.6658 0.3342 0.3342 0.5288   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.666e+00 3.191e-03 522.087 <2e-16 \*\*\*  
## twitter\_data$followers\_count -3.110e-08 1.867e-08 -1.666 0.0958 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4718 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

anova(m\_follower\_gender)

## Analysis of Variance Table  
##   
## Response: as.numeric(twitter\_data$gender)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## twitter\_data$followers\_count 1 0.6 0.61761 2.775 0.09576 .  
## Residuals 21886 4871.0 0.22256   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(m\_follower\_gender)



In the boxplot and density plot, there is no obvious followers count different between female and male. Becasue the gender is binary, I use logistic regression here to measure the relationship between gender and followers count. However, the result of t-test shows that the p-value of slope of followers count is 0.0958, I can't reject the null hyphotheis. The gender and followers have no relationship.

When I observed the residual plot, the error term is not normal and homoscedasticity. And the data seems to have outliers which violate the error term assumptions.

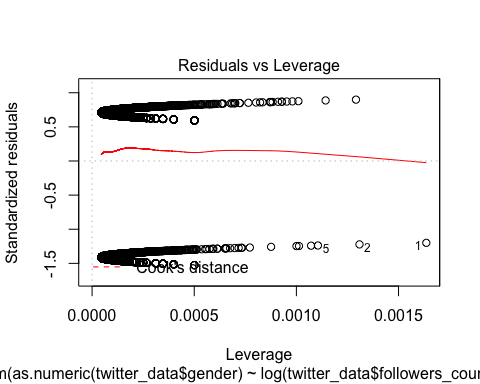
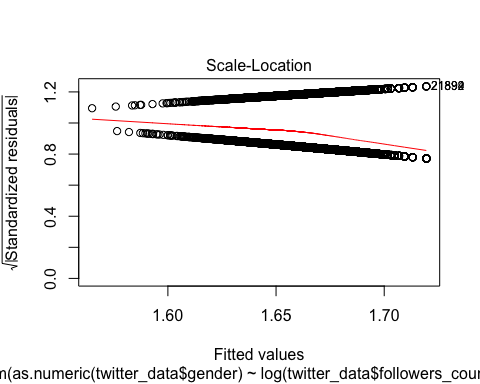
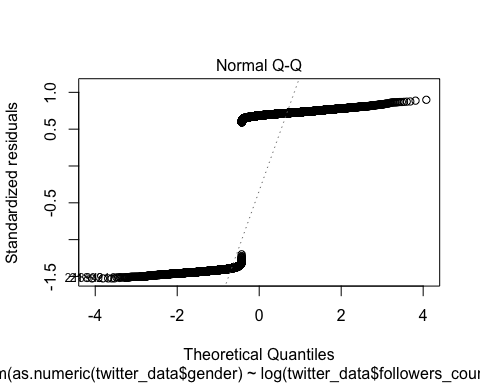
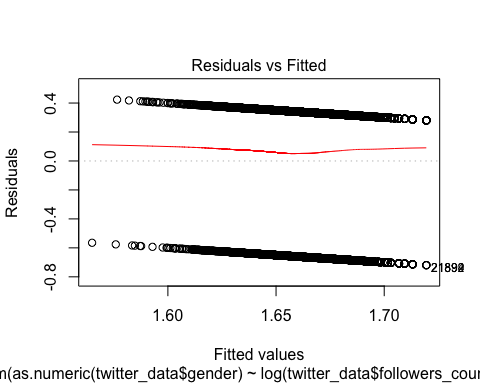
ANOVA test also get the p value which is 0.09576. I can't reject the null hyphothesis. This linear regression is not significant.

I try to log transform the followers\_count data which add 1 pseudo count to original count and observe the relation between gender and followers count

m\_follower\_gender\_log <- lm(as.numeric(twitter\_data$gender)~log(twitter\_data$followers\_count+1))  
  
summary(m\_follower\_gender\_log)

##   
## Call:  
## lm(formula = as.numeric(twitter\_data$gender) ~ log(twitter\_data$followers\_count +   
## 1))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7195 -0.6552 0.3224 0.3380 0.4235   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 1.719458 0.010551 162.970  
## log(twitter\_data$followers\_count + 1) -0.009134 0.001706 -5.353  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## log(twitter\_data$followers\_count + 1) 8.73e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4715 on 21886 degrees of freedom  
## (28 observations deleted due to missingness)  
## Multiple R-squared: 0.001308, Adjusted R-squared: 0.001262   
## F-statistic: 28.66 on 1 and 21886 DF, p-value: 8.729e-08

plot(m\_follower\_gender\_log)



anova(m\_follower\_gender\_log)

## Analysis of Variance Table  
##   
## Response: as.numeric(twitter\_data$gender)  
## Df Sum Sq Mean Sq F value  
## log(twitter\_data$followers\_count + 1) 1 6.4 6.3704 28.657  
## Residuals 21886 4865.3 0.2223   
## Pr(>F)   
## log(twitter\_data$followers\_count + 1) 8.729e-08 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m\_follower\_gender, m\_follower\_gender\_log)

## Analysis of Variance Table  
##   
## Model 1: as.numeric(twitter\_data$gender) ~ twitter\_data$followers\_count  
## Model 2: as.numeric(twitter\_data$gender) ~ log(twitter\_data$followers\_count +   
## 1)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 21886 4871.0   
## 2 21886 4865.3 0 5.7527

The p-value of beta0 and beta1 are small enough to reject the null hyphothesis. However, the error term in residual plot is not normal and homoscedasticity. And the data seems to have outliers which violate the error term assumptions.

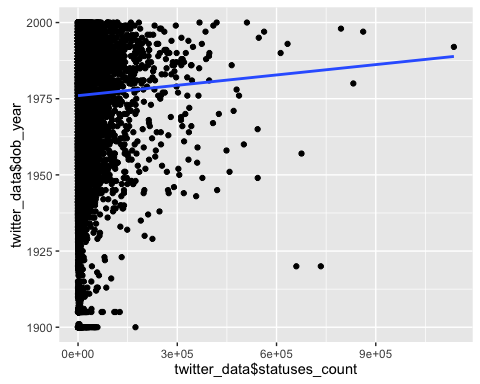
The anova results show that this new linear regression model is significant which the p-value is smaller than 0.001.

Using anova to compare two models, the second model has a little imporvement and better than the firsh model. But both of them can't provide evident of relation between gender and followers count

In sum, both models are not significant and they violated the assumptions

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

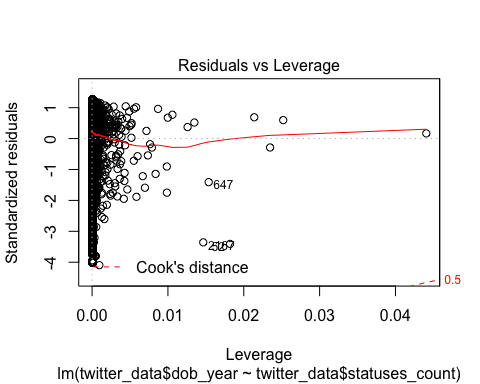
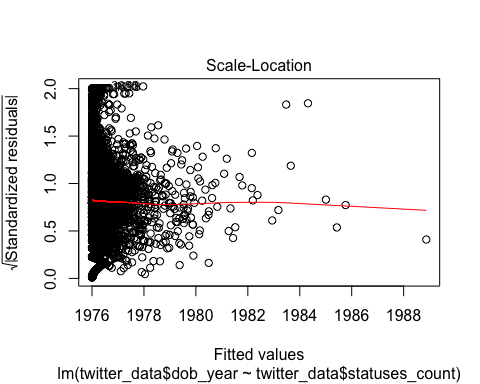
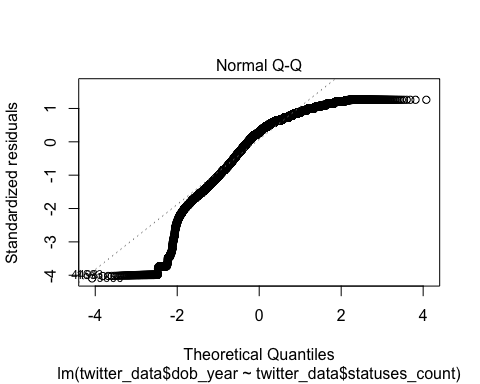
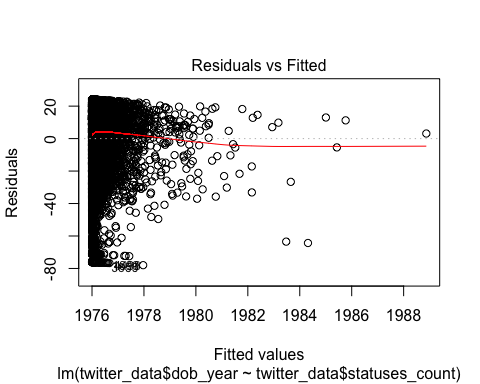
qplot( twitter\_data$statuses\_count, twitter\_data$dob\_year)+ stat\_smooth(method="lm",se=FALSE)



m\_statuses\_dobYear <- lm(twitter\_data$dob\_year ~ twitter\_data$statuses\_count)   
summary(m\_statuses\_dobYear)

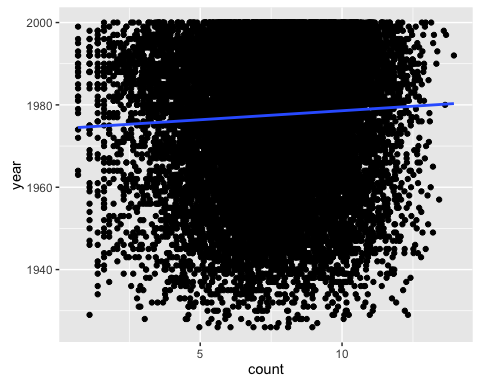
##   
## Call:  
## lm(formula = twitter\_data$dob\_year ~ twitter\_data$statuses\_count)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -77.969 -11.105 5.888 13.983 23.997   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.976e+03 1.361e-01 14516.049 < 2e-16 \*\*\*  
## twitter\_data$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

plot(m\_statuses\_dobYear)



In the plot, I can't find an obvious relation between dob year and statuses count. After linear regression of dob year and statuses count, the p-values of slope are not good enough to explain the data. The value of slope is close to zero. And results of the residual plot is not good either. Because part of the data are much greater than others and the data around year 1900 are not meaningfull which it's little possible that the users' bod year is 1900, I transform the data and trim the outliers.

new\_stat\_count <- log(twitter\_data$statuses\_count+1)  
stat\_bod\_year <- data.frame(count = new\_stat\_count,year = twitter\_data$dob\_year)  
trim\_condition <- stat\_bod\_year$year > 1925  
stat\_bod\_year\_trim <- stat\_bod\_year[trim\_condition,]  
  
qplot(count,year,data=stat\_bod\_year\_trim) + stat\_smooth(method="lm",se=FALSE)

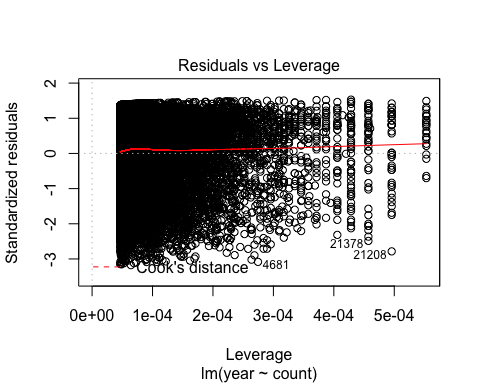
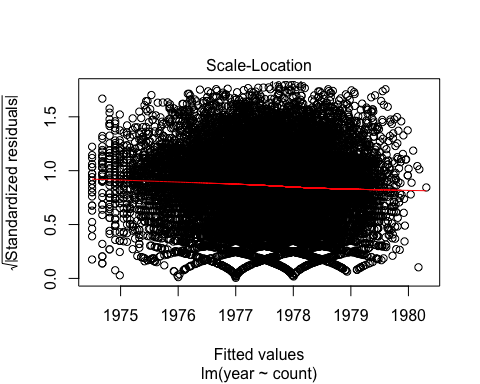
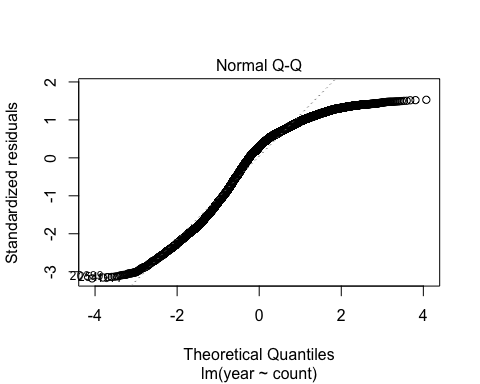
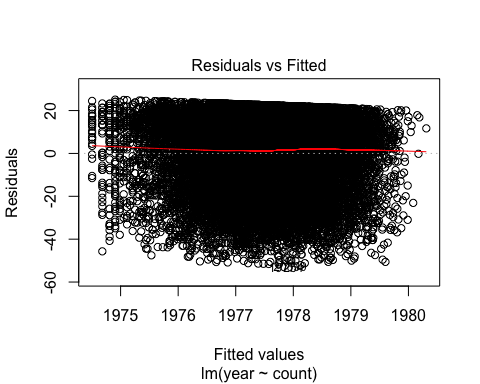


In this new plot, the status count data are in the range of 0 to 20. Now using linear regression to check whether I can get a better model.

m\_log\_stat\_year\_trim <- lm(year~count, data = stat\_bod\_year\_trim)   
summary(m\_log\_stat\_year\_trim)

##   
## Call:  
## lm(formula = year ~ count, data = stat\_bod\_year\_trim)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -52.124 -11.399 4.591 12.803 25.094   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.974e+03 4.212e-01 4686.838 <2e-16 \*\*\*  
## count 4.382e-01 5.306e-02 8.259 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16.41 on 21475 degrees of freedom  
## Multiple R-squared: 0.003166, Adjusted R-squared: 0.00312   
## F-statistic: 68.21 on 1 and 21475 DF, p-value: < 2.2e-16

plot(m\_log\_stat\_year\_trim)



The new linear regression model seems better than the previous one. The p-value of t-test for slope is small enough to reject the null hyphothesis. The residual plot is better than previous one which is not violated the model assumptions.

To compare whethere the transform will provide a better model

stat\_data <- data.frame(count = twitter\_data$statuses\_count, year = twitter\_data$dob\_year )  
trim\_condition <- stat\_bod\_year$year > 1925  
stat\_bod\_year\_trim <- stat\_data[trim\_condition,]  
m\_stat\_year\_trim <- lm(year~count, data = stat\_data)  
m\_log\_stat\_year\_trim <- lm(year~log(count), data = stat\_data)   
  
anova( m\_stat\_year\_trim, m\_log\_stat\_year\_trim)

## Analysis of Variance Table  
##   
## Model 1: year ~ count  
## Model 2: year ~ log(count)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 21914 7952217   
## 2 21914 7934634 0 17583

The result of anova should that the second tranformed data performed better than the first one because of the higher F value.

In summary, the formula of transformed data is

## To find significant linear models

According to the first question, I find in column gender, there are "NA" data. First of all, I remove all NA data from twitter data

trim\_data = subset(twitter\_data, gender =="female" | gender == "male")  
  
#relation of wage and age  
summary(lm(wage~age, data = trim\_data ))

##   
## Call:  
## lm(formula = wage ~ age, data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.012 -9.454 -2.619 5.421 82.002   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.294e+01 2.920e-01 78.58 <2e-16 \*\*\*  
## age 8.536e-04 7.732e-03 0.11 0.912   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.61 on 21886 degrees of freedom  
## Multiple R-squared: 5.568e-07, Adjusted R-squared: -4.513e-05   
## F-statistic: 0.01219 on 1 and 21886 DF, p-value: 0.9121

#realation of wage and education  
summary(lm(wage~education, data = trim\_data ))

##   
## Call:  
## lm(formula = wage ~ education, data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.355 -9.441 -2.618 5.464 81.998   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.24325 0.47775 46.558 <2e-16 \*\*\*  
## education 0.05851 0.03740 1.565 0.118   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.6 on 21886 degrees of freedom  
## Multiple R-squared: 0.0001118, Adjusted R-squared: 6.614e-05   
## F-statistic: 2.448 on 1 and 21886 DF, p-value: 0.1177

#realation of wage and height  
summary(lm(wage~height,data = trim\_data ))

##   
## Call:  
## lm(formula = wage ~ height, data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.039 -9.126 -2.350 5.415 83.718   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.02944 1.91308 -13.61 <2e-16 \*\*\*  
## height 0.28533 0.01112 25.65 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.39 on 21886 degrees of freedom  
## Multiple R-squared: 0.02918, Adjusted R-squared: 0.02914   
## F-statistic: 657.8 on 1 and 21886 DF, p-value: < 2.2e-16

#relation of wage and experience  
summary(lm(wage~experience,data = trim\_data))

##   
## Call:  
## lm(formula = wage ~ experience, data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.189 -9.453 -2.601 5.425 81.990   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.081355 0.129419 178.346 <2e-16 \*\*\*  
## experience -0.009817 0.007694 -1.276 0.202   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.6 on 21886 degrees of freedom  
## Multiple R-squared: 7.438e-05, Adjusted R-squared: 2.869e-05   
## F-statistic: 1.628 on 1 and 21886 DF, p-value: 0.202

#relation of age and followers\_count  
summary(lm(age~log(followers\_count+1),data = trim\_data))

##   
## Call:  
## lm(formula = age ~ log(followers\_count + 1), data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.579 -7.568 0.447 8.464 55.431   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.70086 0.28572 124.949 <2e-16 \*\*\*  
## log(followers\_count + 1) -0.02741 0.04620 -0.593 0.553   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.77 on 21886 degrees of freedom  
## Multiple R-squared: 1.608e-05, Adjusted R-squared: -2.961e-05   
## F-statistic: 0.3519 on 1 and 21886 DF, p-value: 0.5531

#relation of education and followers\_count  
summary(lm(education~log(followers\_count+1),data = trim\_data))

##   
## Call:  
## lm(formula = education ~ log(followers\_count + 1), data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.4984 -1.4984 0.5008 1.5021 11.5016   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.4992934 0.0590654 211.618 <2e-16 \*\*\*  
## log(followers\_count + 1) -0.0001979 0.0095516 -0.021 0.983   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.639 on 21886 degrees of freedom  
## Multiple R-squared: 1.961e-08, Adjusted R-squared: -4.567e-05   
## F-statistic: 0.0004292 on 1 and 21886 DF, p-value: 0.9835

#relation of heigh and followers\_count  
summary(lm(height~log(followers\_count+1),data = trim\_data))

##   
## Call:  
## lm(formula = height ~ log(followers\_count + 1), data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.173 -7.149 0.443 6.721 31.750   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 172.66988 0.19556 882.931 < 2e-16 \*\*\*  
## log(followers\_count + 1) -0.15650 0.03163 -4.949 7.53e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.739 on 21886 degrees of freedom  
## Multiple R-squared: 0.001118, Adjusted R-squared: 0.001072   
## F-statistic: 24.49 on 1 and 21886 DF, p-value: 7.529e-07

The results find two significant linear models. One is wage and height which the p-values of beta0 and beta1 is smaller than 0.001. And the other one is height and log transformed followers\_count.

### linear model of wage and height

m\_wage\_height <- lm (wage~height, data=trim\_data)  
summary(m\_wage\_height)

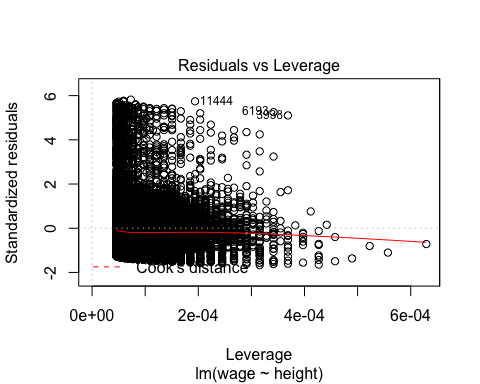
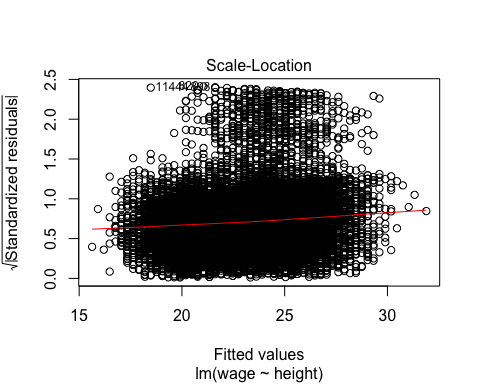
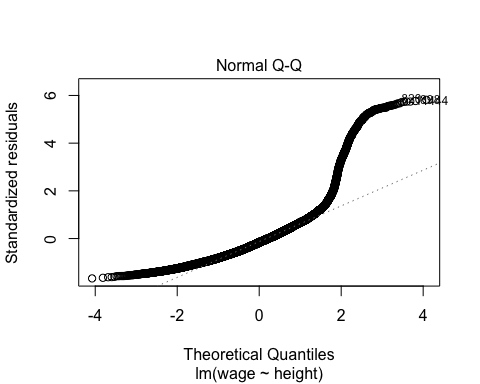
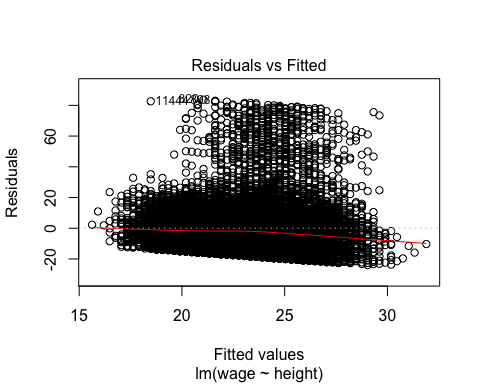
##   
## Call:  
## lm(formula = wage ~ height, data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.039 -9.126 -2.350 5.415 83.718   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.02944 1.91308 -13.61 <2e-16 \*\*\*  
## height 0.28533 0.01112 25.65 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.39 on 21886 degrees of freedom  
## Multiple R-squared: 0.02918, Adjusted R-squared: 0.02914   
## F-statistic: 657.8 on 1 and 21886 DF, p-value: < 2.2e-16

* Is the relationship significant?

According to results of t-test, it shows that the p-value of intercept and solpe are all smaller than 0.001, which means I can reject the null hyphothesis and accept the althernative one. The relationship is significant. The model is:

&& {wage} = -26.02944 + 0.28533\* x\_{height} + &&

plot(m\_wage\_height)



* Are any model assumptions violated?

In the residual vs fitted plot, the residuals are not bounce randomly arount the 0 line. It should be evenly distributed on the both side of the 0 line. Here, the residual are mainly located above the horizon line. So the residuals are not homoscedastic. It violated the model assumptions.

In the QQ plot, only the data between -1 to 1 are on the dashed line. most of the dots are not on the dashed line. If the residuals are normal distribution, most of the dot should be on the dashed line. So the residuals are not from noraml distribution. It violated the model assumptions.

In the Scale Location plot, It is easy to find a similar pattern as residual vs fitted plot. This graph also show whether the residuals are homoscedastic. Most of the data are above the red line. And the red line is not flat and is skewed. So the residuals are not homoscedastic. It violated the model assumptions.

In the residuals vs leverage plot, the red smoothed line stays close to the horizontal gray dashed line, and no points have a large Cook's distance. So there are no points which will influence the regression significantly.

* Does the model make sense?

The linear regression shows that the height and wage have a positive relationship. If the height increase 1 unit, the wage will increase 0.285 unit. However, The relationship of height and wage reflect the relationship of gender and wage. In last lesson, we know that men make more money than woman, and men is taller than women. Now, I decide to use only women' wage and height data to see whether they still have this relationship

trim\_data\_F <- subset(twitter\_data, gender =="female" )  
  
m\_wage\_height\_F <- lm (wage~height, data=trim\_data\_F)  
summary(m\_wage\_height\_F)

##   
## Call:  
## lm(formula = wage ~ height, data = trim\_data\_F)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.482 -7.050 -0.853 5.618 35.663   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 19.81937 3.83785 5.164 2.48e-07 \*\*\*  
## height -0.00886 0.02362 -0.375 0.708   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.648 on 7317 degrees of freedom  
## Multiple R-squared: 1.923e-05, Adjusted R-squared: -0.0001174   
## F-statistic: 0.1407 on 1 and 7317 DF, p-value: 0.7076

The P-value of solpe in t-test is 0.708 which is larger then 0.001. I can't reject the null hyphothesis. That means the linear model of wage and height is not longer significant with the control of gender.

The height is associated with gender so this model makes sense.

### linear model of height and log transformed followers\_count

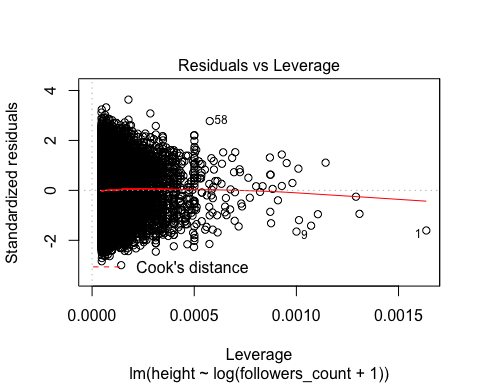
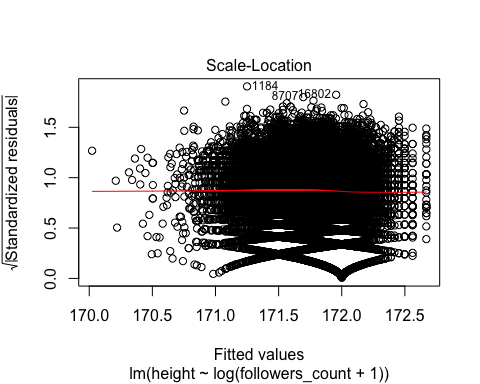
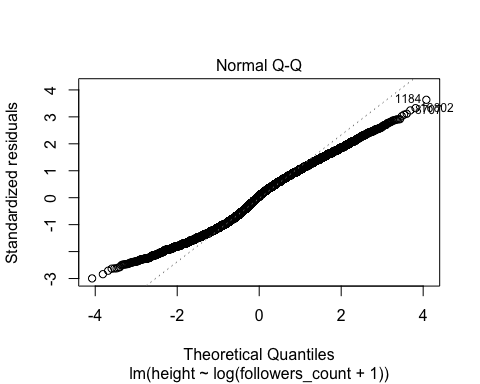
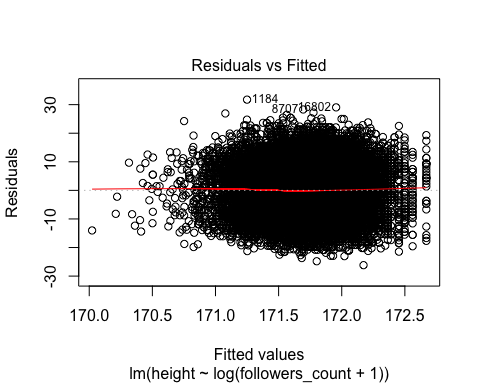
m\_height\_followers <- lm (height~log(followers\_count+1),data = trim\_data)  
summary(m\_height\_followers)

##   
## Call:  
## lm(formula = height ~ log(followers\_count + 1), data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.173 -7.149 0.443 6.721 31.750   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 172.66988 0.19556 882.931 < 2e-16 \*\*\*  
## log(followers\_count + 1) -0.15650 0.03163 -4.949 7.53e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.739 on 21886 degrees of freedom  
## Multiple R-squared: 0.001118, Adjusted R-squared: 0.001072   
## F-statistic: 24.49 on 1 and 21886 DF, p-value: 7.529e-07

* Is the relationship significant?

According to results of t-test, it shows that the p-value of intercept and solpe are all smaller than 0.001, which means I can reject the null hyphothesis and accept the althernative one. The relationship is significant. The model is:

plot(m\_height\_followers)



* Are any model assumptions violated?

In the residual vs fitted plot, the residuals are bounce randomly arount the 0 line. It shows evenly distributed on the both side of the 0 line. So the residuals are homoscedastic. It didn't violate the model assumptions.

In the QQ plot, only the data between -1 to 1 are on the dashed line. most of the dots are not on the dashed line. If the residuals are normal distribution, most of the dot should be on the dashed line. So the residuals are not from noraml distribution. It violated the model assumptions.

In the Scale Location plot, It shows a similar pattern as residual vs fitted plot. Most of the data are on the both sides of the red line. And the red line is flat. So the residuals are homoscedastic. It didn't violate the model assumptions.

In the residuals vs leverage plot, the red smoothed line stays close to the horizontal gray dashed line, and no points have a large Cook's distance. So there are no points which will influence the regression significantly.

* Does the model make sense? The linear regression shows that the height and followers count have a negtive relationship. If the log transtform followers count increase 1 unit, the height will decrease 0.1565 unit. This model is useful for twitter to predict the height of the users.Howeve, we have known that the height may be effected by gender according to the last model. Now, I decide to control the gender variable and test whether they still have this relationship

trim\_data\_F <- subset(twitter\_data, gender =="female" )  
  
m\_height\_followers\_F <- lm (height~log(followers\_count+1), data=trim\_data\_F)  
summary(m\_height\_followers\_F)

##   
## Call:  
## lm(formula = height ~ log(followers\_count + 1), data = trim\_data\_F)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.5141 -2.5274 -0.3063 2.6111 16.5208   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 162.60434 0.16934 960.203 <2e-16 \*\*\*  
## log(followers\_count + 1) -0.02840 0.02701 -1.052 0.293   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.28 on 7317 degrees of freedom  
## Multiple R-squared: 0.0001511, Adjusted R-squared: 1.446e-05   
## F-statistic: 1.106 on 1 and 7317 DF, p-value: 0.293

The P-value of solpe in t-test is 0.293 which is larger then 0.001. I can't reject the null hyphothesis. That means the linear model of log transformed followers count and height is not longer significant with the control of gender.

It seems women have more followers than men. and this model makes sense.

## A multivariate relation between wage & height, race, age, education, and experience

First, I must define response variable and explanatory variables. For these variables, wage is better to choose as explanatory variable and the others are used as explanatory.

* Is the relationship significant?

m\_multi <- lm(wage~height+race+age+education+experience, data = trim\_data)  
summary(m\_multi)

##   
## Call:  
## lm(formula = wage ~ height + race + age + education + experience,   
## data = trim\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.779 -9.119 -2.352 5.371 83.892   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -27.904320 2.275385 -12.264 <2e-16 \*\*\*  
## height 0.285271 0.011169 25.541 <2e-16 \*\*\*  
## raceasian 0.978937 1.152451 0.849 0.3956   
## racehispanic 1.817758 1.302891 1.395 0.1630   
## raceindian 1.030086 1.545759 0.666 0.5052   
## racelatino 1.114809 1.139324 0.978 0.3278   
## racemixed 0.389607 1.467286 0.266 0.7906   
## racenative american 0.637755 1.385751 0.460 0.6454   
## racepacific islander 0.429773 1.365592 0.315 0.7530   
## racepersian 1.079600 1.290306 0.837 0.4028   
## racewhite 0.892069 1.059539 0.842 0.3998   
## age 0.006231 0.007626 0.817 0.4139   
## education 0.063426 0.036865 1.720 0.0854 .   
## experience -0.003230 0.007608 -0.425 0.6712   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.39 on 21874 degrees of freedom  
## Multiple R-squared: 0.02949, Adjusted R-squared: 0.02891   
## F-statistic: 51.12 on 13 and 21874 DF, p-value: < 2.2e-16

anova(m\_multi)

## Analysis of Variance Table  
##   
## Response: wage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## height 1 136235 136235 657.6921 < 2e-16 \*\*\*  
## race 9 645 72 0.3461 0.95957   
## age 1 138 138 0.6647 0.41490   
## education 1 615 615 2.9698 0.08485 .   
## experience 1 37 37 0.1802 0.67120   
## Residuals 21874 4531003 207   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

According to the results of anova, the p-value of height in F-test is smaller than 0.001, so I can reject the null hyphothesis which all betas are equal to zero. The realtionship is significant.

* Are any model assumptions violated?

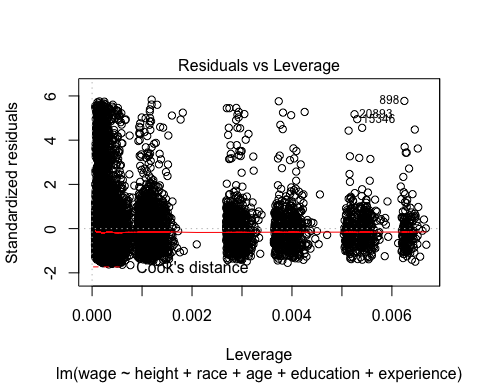
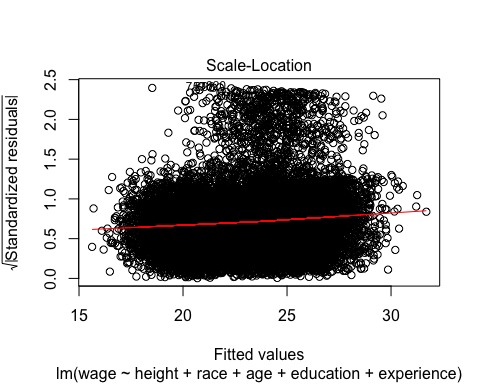
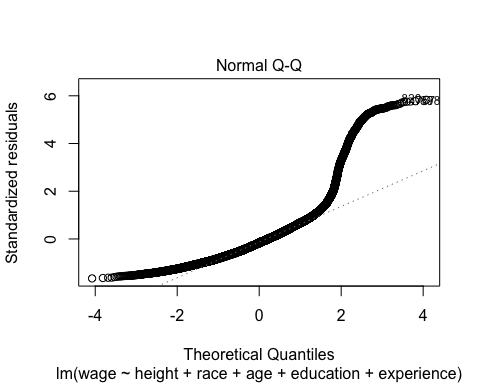
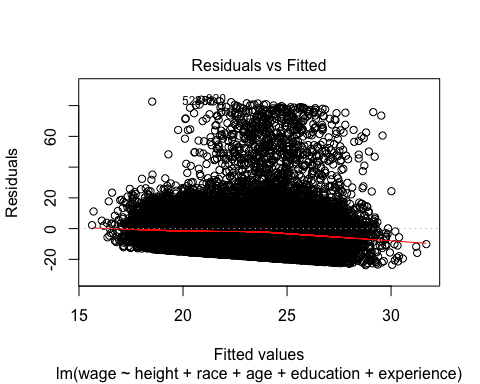
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

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

library(nortest)  
plot(m\_multi)



model\_resid <- resid(m\_multi)  
  
#test normality using Lilliefors (Kolmogorov-Smirnov) normality test  
lillie.test(model\_resid)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: model\_resid  
## D = 0.10457, p-value < 2.2e-16

#test the homoscedasticity using Breusch and Pagan Test  
bptest(m\_multi, studentize = FALSE )

##   
## Breusch-Pagan test  
##   
## data: m\_multi  
## BP = 1354.2, df = 13, p-value < 2.2e-16

In the residual vs fitted plot, the residuals are not bounce randomly arount the 0 line. It should be evenly distributed on the both side of the 0 line. Here, the residual are mainly located above the horizon line. I also use Breusch and Pagan test to check the homoscedastic. The p-value of test is smaller than 0.001, and I can reject the null hyphothesis. So the residuals are not homoscedastic. It violated the model assumptions.

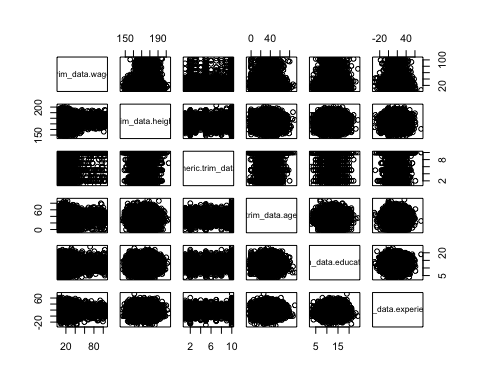
In the QQ plot, only the data between -1 to 2 are on the dashed line. most of the dots are not on the dashed line. If the residuals are normal distribution, most of the dot should be on the dashed line. I also use Lilliefors (Kolmogorov-Smirnov) normality test to check the normality. The p-value is smaller than 0.001, and I can reject the null hyphothesis. So the residuals are not from noraml distribution. It violated the model assumptions.

In the Scale Location plot, It is easy to find a similar pattern as residual vs fitted plot. This graph also show whether the residuals are homoscedastic. Most of the data are above the red line. And the red line is not flat and is skewed. So the residuals are not homoscedastic. It violated the model assumptions.

In the residuals vs leverage plot, the red smoothed line stays close to the horizontal gray dashed line, and no points have a large Cook's distance. So there are no points which will influence the regression significantly.

* Is there any multi-colinearity in multivariate models?

trim\_data\_mul <- data.frame(trim\_data$wage, trim\_data$height, as.numeric(trim\_data$race), trim\_data$age, trim\_data$education, trim\_data$experience )  
  
pairs(trim\_data\_mul)



names(trim\_data\_mul) <- c("wage","height","race","age", "education", "experience")  
  
cor(trim\_data\_mul)

## wage height race age  
## wage 1.0000000000 0.170823504 -0.0025550804 0.0007461989  
## height 0.1708235045 1.000000000 -0.0010195519 -0.0272166124  
## race -0.0025550804 -0.001019552 1.0000000000 0.0049404977  
## age 0.0007461989 -0.027216612 0.0049404977 1.0000000000  
## education 0.0105747252 -0.004711350 -0.0003826532 0.0028553476  
## experience -0.0086244950 -0.032473770 -0.0058896290 0.0179960099  
## education experience  
## wage 0.0105747252 -0.008624495  
## height -0.0047113500 -0.032473770  
## race -0.0003826532 -0.005889629  
## age 0.0028553476 0.017996010  
## education 1.0000000000 -0.006489882  
## experience -0.0064898817 1.000000000

Calculating the correlation of each two variables, it shows that no two predictor variables in this model are highly correlated. No multi-colinearity is in this model.

* In in multivariate models are predictor variables independent of all the other predictor variables?

for (i in (2:5)){  
 for (j in ((i+1):6)){  
 test <- cor.test(trim\_data\_mul[,i],trim\_data\_mul[,j])  
 print(c(colnames(trim\_data\_mul)[i], colnames(trim\_data\_mul)[j]))  
 print(test$p.value)  
 }  
}

## [1] "height" "race"   
## [1] 0.8801098  
## [1] "height" "age"   
## [1] 5.646826e-05  
## [1] "height" "education"  
## [1] 0.4858096  
## [1] "height" "experience"  
## [1] 1.54477e-06  
## [1] "race" "age"   
## [1] 0.4648468  
## [1] "race" "education"  
## [1] 0.9548569  
## [1] "race" "experience"  
## [1] 0.3835879  
## [1] "age" "education"  
## [1] 0.672723  
## [1] "age" "experience"  
## [1] 0.007756255  
## [1] "education" "experience"  
## [1] 0.3370014

According to the results of correlation test, the height vs age, height vs experience have non-zero correlation which means this two groups are not indepent, becauset the p-value of correlation test of them are both smaller than 0.001 and I can reject the null hypothesis. the height vs age and the height vs experience are not independent.

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

Using backward elimination to select the best predictors.

beg <- lm(wage~height+race+age+education+experience, data=trim\_data\_mul)  
end <- lm(wage~., data=trim\_data\_mul)  
empty <- lm(wage ~ 1, data = trim\_data\_mul)  
bounds <- list(upper = end, lower = empty)  
backward\_elim\_reg <- step(beg, bounds, direction = "backward")

## Start: AIC=116738.5  
## wage ~ height + race + age + education + experience  
##   
## Df Sum of Sq RSS AIC  
## - race 1 27 4531654 116737  
## - experience 1 45 4531672 116737  
## - age 1 138 4531764 116737  
## <none> 4531627 116738  
## - education 1 601 4532227 116739  
## - height 1 136147 4667774 117384  
##   
## Step: AIC=116736.6  
## wage ~ height + age + education + experience  
##   
## Df Sum of Sq RSS AIC  
## - experience 1 45 4531699 116735  
## - age 1 137 4531791 116735  
## <none> 4531654 116737  
## - education 1 601 4532255 116737  
## - height 1 136151 4667806 117383  
##   
## Step: AIC=116734.8  
## wage ~ height + age + education  
##   
## Df Sum of Sq RSS AIC  
## - age 1 134 4531833 116733  
## <none> 4531699 116735  
## - education 1 603 4532302 116736  
## - height 1 136450 4668148 117382  
##   
## Step: AIC=116733.5  
## wage ~ height + education  
##   
## Df Sum of Sq RSS AIC  
## <none> 4531833 116733  
## - education 1 605 4532438 116734  
## - height 1 136318 4668151 117380

backward\_elim\_reg

##   
## Call:  
## lm(formula = wage ~ height + education, data = trim\_data\_mul)  
##   
## Coefficients:  
## (Intercept) height education   
## -26.83182 0.28542 0.06297

According to the backward elimination regression results, The smaller AIC is, the better result. So I choose two most significant predictor variables and get the model:

To check whether this model have interaction terms

m\_mul\_wage\_final <- lm(wage ~ height+education+height:education , data= trim\_data\_mul)  
  
anova(m\_mul\_wage\_final)

## Analysis of Variance Table  
##   
## Response: wage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## height 1 136235 136235 657.8972 < 2e-16 \*\*\*  
## education 1 605 605 2.9196 0.08752 .   
## height:education 1 173 173 0.8347 0.36094   
## Residuals 21884 4531661 207   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The p-value of F-test of interaction term is 0.36 which is larger then 0.01. I can't reject the null hyphothesis. There is no interaction term.

To analyze the final two variables

m\_mul\_wage\_final <- lm(wage ~ height+education , data= trim\_data\_mul)  
summary(m\_mul\_wage\_final )

##   
## Call:  
## lm(formula = wage ~ height + education, data = trim\_data\_mul)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.072 -9.137 -2.349 5.393 83.939   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.83182 1.96978 -13.622 <2e-16 \*\*\*  
## height 0.28542 0.01112 25.657 <2e-16 \*\*\*  
## education 0.06297 0.03685 1.709 0.0875 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.39 on 21885 degrees of freedom  
## Multiple R-squared: 0.02931, Adjusted R-squared: 0.02922   
## F-statistic: 330.4 on 2 and 21885 DF, p-value: < 2.2e-16

anova(m\_mul\_wage\_final )

## Analysis of Variance Table  
##   
## Response: wage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## height 1 136235 136235 657.9022 < 2e-16 \*\*\*  
## education 1 605 605 2.9196 0.08752 .   
## Residuals 21885 4531833 207   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

T-test results show that the p-value of education is 0.08752 which is larger than 0.001, and I can't reject the null hyphothesis. The education variable is not significant. And the F-test results show the the p-value of height is small enough to reject the null hyphotheis, this model is significant.

So the multiple linear regression model fomula is

* Does the model make sense?

In my final model, the wage has a positive relationship with height and education. I thinks it makes sense.

First, according to the previous analysis, height is associate with gender. In genera, men is taller than women, and men earn more than women. The relationship between wage and gender refect the relationship between gender and wage.

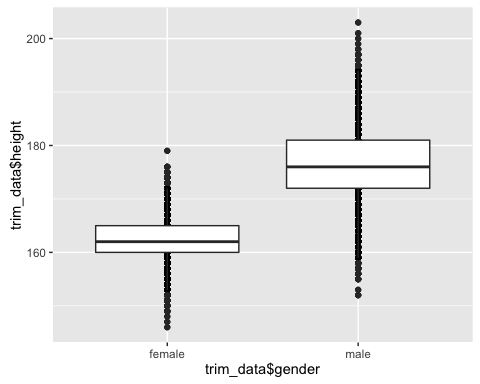
Second, my model also shows that the higher education a person is, the more he/she earns. The higher education means take more time to improve herself/himself. The oppotunity cost will be paid by higher wage.

Third, there is no interaction of height and education also explain the fact the getting a high education has nothing to do with height. And the coefficient of each variable show the extent that it effect on wage which gender(height) have more effect on wage than education.

### A significant logistic linear model of your choosing

In gender, there are "NA" data. I will not choose these data.

trim\_data = subset(twitter\_data, gender =="female" | gender == "male")  
qplot( trim\_data$gender, trim\_data$height) + geom\_boxplot()



There are no obvious outliers in the boxplot, the trimmed data can be used in following analysis.

* Is the relationship significant?

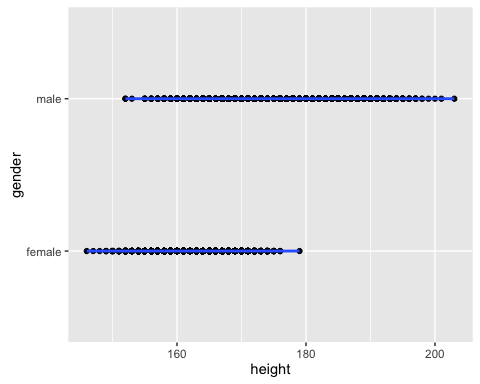
m\_height\_gender <- lm(as.numeric(trim\_data$gender)~trim\_data$height)  
summary(m\_height\_gender)

##   
## Call:  
## lm(formula = as.numeric(trim\_data$gender) ~ trim\_data$height)  
##   
## 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 \*\*\*  
## trim\_data$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  
## 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

anova(m\_height\_gender)

## Analysis of Variance Table  
##   
## Response: as.numeric(trim\_data$gender)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## trim\_data$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

qplot(height, gender, data = trim\_data )+stat\_smooth(method="lm",se=FALSE)



The p-value of t-test of beta0 and beta1 are smaller than 0.001, and I can reject the null hyphothesis. These two coefficients are significant.

The p-value of F-test is also smaller than 0.001, and I can reject the null hyphothesis. This model is singinicant.

This logistic linear model's formula is:

* Are any model assumptions violated?

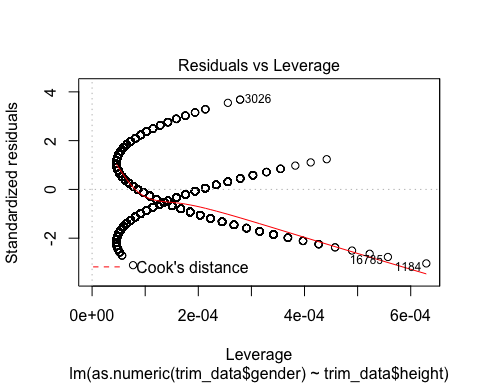
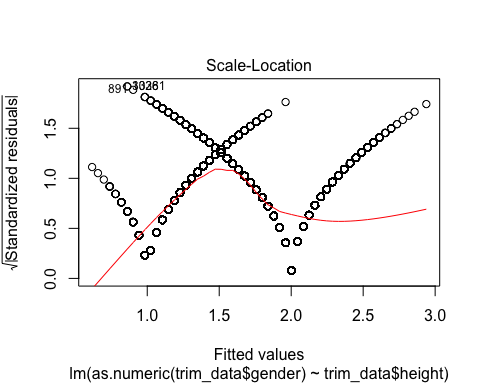
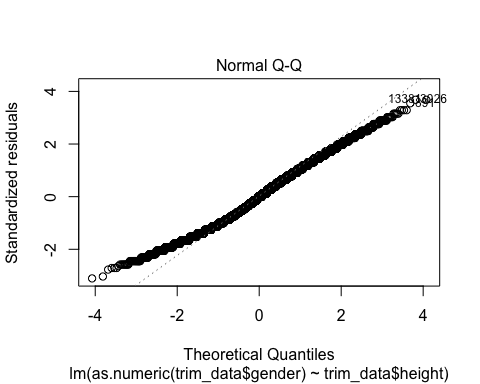
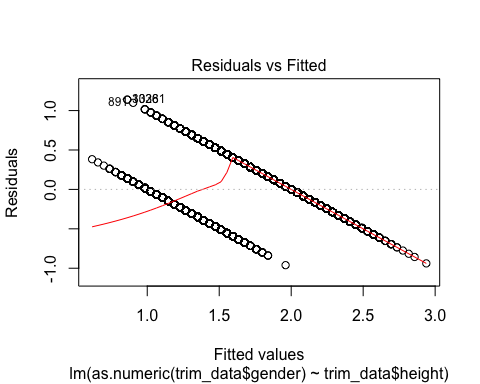
bptest(m\_height\_gender, studentize = F)

##   
## Breusch-Pagan test  
##   
## data: m\_height\_gender  
## BP = 987.53, df = 1, p-value < 2.2e-16

lillie.test(resid(m\_height\_gender))

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: resid(m\_height\_gender)  
## D = 0.045274, p-value < 2.2e-16

plot(m\_height\_gender)



I use Breusch-Pagan test to check the homoscedastic. The p-value is smaller than 0.001 which I reject the null hyphothesis. The residuals are not homoscedastic. The residuals plot - Residuals vs Fitted and Scale-location also prove it becasue the red lines of both plots are not flat.

I use Lilliefors (Kolmogorov-Smirnov) normality test to check the normality. The p-value is smaller than 0.001 which I reject the null hyphothesis. The residuals are not from normal distribution. The QQ-plot prove it too. Only the dot from range -1 to 1 are on the dashed line and most of the dots are not on the line.

Although there are no significant leverage dots in the data, but red line is not flat. No points have a large Cook's distance. In all, there are no points which will influence the regression significantly.

In summary, the residuals violated assumptions.

* Does the model make sense? Why or why not?

The model makes sense. In this model, height and gender have a positive relationship. This heighter a person is, the more possible the person is man. And it is close to the fact that men is taller than woman on average.