Module\_02 Solution

Romil Shah

05-31-2016

### Load twitter data and needed additional packages

require(ggplot2)

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.2.5

require(reshape2)

## Loading required package: reshape2

## Warning: package 'reshape2' was built under R version 3.2.5

require(car)

## Loading required package: car

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,  
## logical.return = TRUE, : there is no package called 'car'

require(lmtest)

## Loading required package: lmtest

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,  
## logical.return = TRUE, : there is no package called 'lmtest'

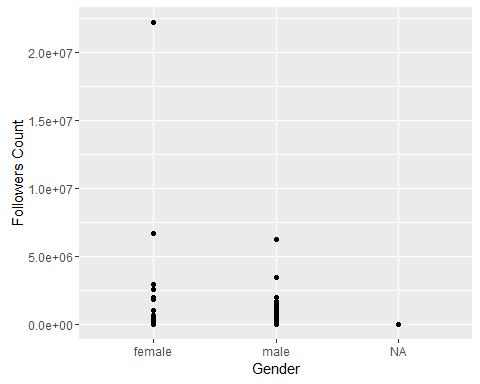
twitter<-read.csv("C:/Users/rams1/Desktop/DSCS6030/Module\_02/M01\_quasi\_twitter.csv")

### 1. A Relation Between followers\_count & gender:

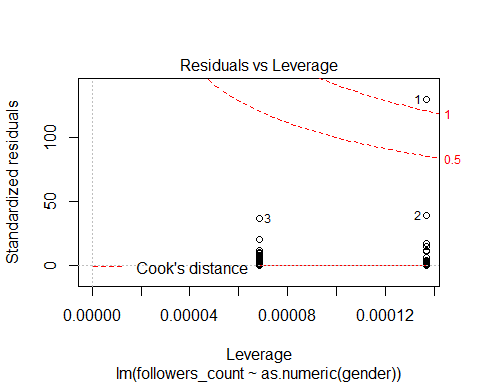
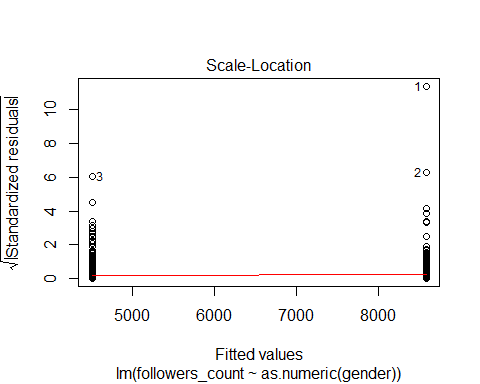
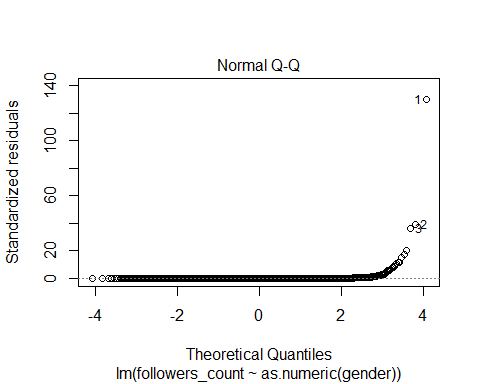
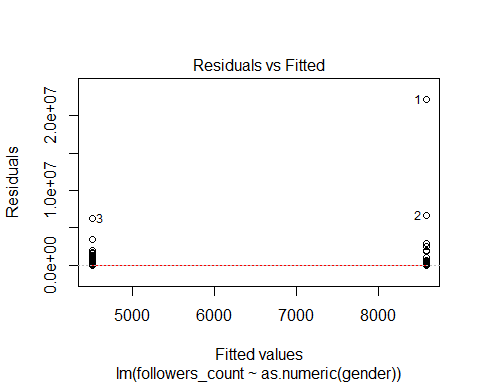
lmfromRMD <- lm(followers\_count ~ as.numeric(gender), data=twitter)  
summary(lmfromRMD)

##   
## Call:  
## lm(formula = followers\_count ~ as.numeric(gender), data = twitter)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8580 -7443 -4373 -3823 22179063   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12656 4237 2.987 0.00282 \*\*  
## as.numeric(gender) -4077 2447 -1.666 0.09576 .   
## ---  
## 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

qplot(gender, followers\_count,data=twitter) + geom\_smooth(method=lm,se=FALSE) + ylab("Followers Count") + xlab("Gender")

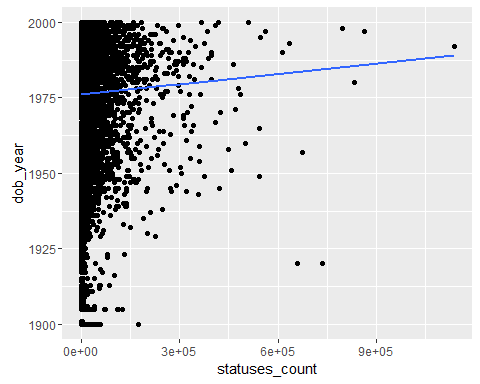


plot(lmfromRMD)

 The relationship is not significant and is violated. This is because gender is a binary variable, thus a logistic regression model would be used. This relation has no connection.

### 2. A Relation between dob\_year & statuses\_count:

qplot(statuses\_count, dob\_year, data=twitter) +geom\_smooth(method=lm, se=FALSE)



lm\_scount\_dobyear1<-lm(statuses\_count~ dob\_year, data=twitter)  
lm\_scount\_dobyear2<-lm(dob\_year ~ statuses\_count, data=twitter)  
summary(lm\_scount\_dobyear1)

##   
## Call:  
## lm(formula = statuses\_count ~ dob\_year, data = twitter)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13456 -11818 -9965 -3135 1123065   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -68218.30 25342.94 -2.692 0.00711 \*\*  
## dob\_year 40.84 12.82 3.185 0.00145 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36170 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

summary(lm\_scount\_dobyear2)

##   
## Call:  
## lm(formula = dob\_year ~ statuses\_count, data = twitter)  
##   
## 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

Both do not look good, with either poor p-values or almost 0 slopes. Thus the relation is not siginficant. Here as the data is skewed transformation can provide some relationship.

### A significant linear model of your choosing:

Test the relationship between wage and height

new <- na.omit(twitter)  
new $ gender\_F [new $ gender == "female"] <- 1  
new $ gender\_F [new $ gender == "male"] <- 0  
  
head(new[,c("gender", "gender\_F", "friends\_count")])

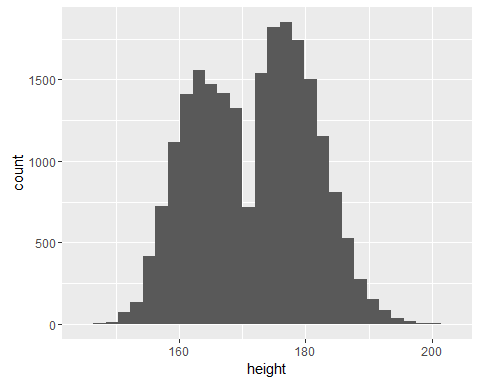
## gender gender\_F friends\_count  
## 1 female 1 1087  
## 2 female 1 5210  
## 3 male 0 1015  
## 4 male 0 338  
## 5 female 1 641  
## 6 female 1 917

summary(new$height)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 146.0 165.0 172.0 171.7 178.0 203.0

qplot(height, data=new)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Proceeding with regression analysis to study more about the relationship. Height predictor variable: (x) Wage outcome variable: (y)

lm.w <- lm(wage ~ height, data=new)  
summary(lm.w)

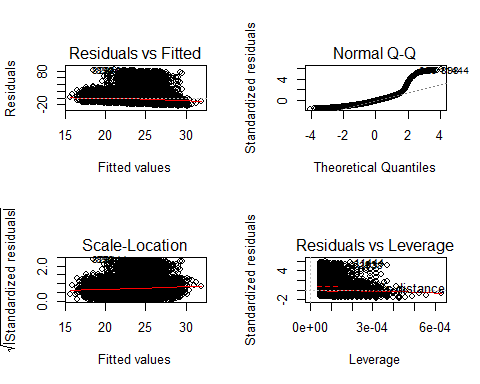
##   
## Call:  
## lm(formula = wage ~ height, data = new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.039 -9.126 -2.350 5.415 83.718   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.02944 1.91308 -13.61 <2e-16 \*\*\*  
## height 0.28533 0.01112 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

1. Is the relationship significant?

1 unit increase in height corresponds to 0.285 increase in wage. The p-value for the beta coefficient for height variable is less than 0.001 at 0.001 level of significance. Thus alternate hypothesis is considered. And thus the relationship is signinificant.

1. Are any model assumptions violated?

par(mfrow=c(2,2))  
plot(lm.w)



\*The residuals vs fitted plot shows that the residuals are not evenly distributed on both sides of the horizonal line with higher concentration of residuals scattered above zero. Thus the constant variance might be violated.

\*Normal-QQ plot shows that the residuals drift away from the dash line towards the end, which indicates a skewed distribution. Therefore, normality is violated.

\*The Scale-Location plot shows a violation of homoscedasticity.

\*The residuals vs leverage plot shows that the red smooth line stays close to the horizontal gray dashed line and there are no obvious influential points.

1. Is there any multi-colinearity in multivariate models?

df <- data.frame(a = twitterrace), h = twittereducation, ex = twitter$experience) cor(df)

Thus it is seen that in this case there does not exist a multi-colinearity in multivariate models.

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

In a multivariate model, the dependence or independence is guided by the model test through regression. The variables whose effect is assumed to exist on others are called independent variables and the ones which have the effect are dependent. But mostly the dependent variables could well serve as independent variables in multivariate analysis.

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

As per the analysis, the wage and height have a good prediction ability.

1. Does the model make sense?

The regression analysis suggests a positive linear relationship between height and wage, i.e. with more height the person is linkely to earn more. THe gender variable too might be acting as an indirect relation. As men tend to be taller than women thus the wages are higher.