Linear Regression vs Generlized Linear Regression Model

Yan Lin

December 10, 2017

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(stringr)


data <- read.csv("C:/Users/Yan/Documents/auto.csv",header=TRUE,",",stringsAsFactors = FALSE)
data <- na.omit(data)
##9134 records and 26 variables

##get rid of the columns of "response", "State" and "Customer" and change the value in effective to dat
data$Effective.To.Date <- as.Date(data$Effective.To.Date,format="%m/%d/%Y")

##calculate the date of accident
data$accidentDate <- data$Effective.To.Date + data$Months.Since.Policy.Inception * 365/12


##claim range
data$Claim.Range <- c("<100","100-200","200-300","300-400","400-500","500-1000","1000-2000","2000-3000")
  findInterval(data$Total.Claim.Amount,c(-Inf,100.5,200.5,300.5,400.5,500.5,1000.5,2000.5,Inf))
]




##Training data
train.data <- data %>%
  filter(accidentDate < "2018-01-01") %>%
  select(Total.Claim.Amount,Claim.Range,Claim.Reason,EmploymentStatus,Location.Code,Gender,Number.of.Op

head(train.data)
```

```
##   Total.Claim.Amount Claim.Range Claim.Reason EmploymentStatus
## 1           384.8111     300-400    Collision         Employed
## 2          1131.4649   1000-2000 Scratch/Dent       Unemployed
## 3           566.4722    500-1000    Collision         Employed
```

```
## 4           529.8813    500-1000    Collision       Unemployed
## 5           138.1309     100-200    Collision         Employed
## 6           321.6000     300-400    Collision         Employed
##   Location.Code Gender Number.of.Open.Complaints
## 1      Suburban      F                         0
## 2      Suburban      F                         0
## 3      Suburban      F                         0
## 4      Suburban      M                         0
## 5         Rural      M                         0
## 6      Suburban      F                         0
```

```
##Test data
test.data <- data %>%
  filter(accidentDate >= "2018-01-01") %>%
  select(Total.Claim.Amount,Claim.Range,Claim.Reason,EmploymentStatus,Location.Code,Gender,Number.of.Ope

##severity analysis
##build a linear regression model from trainning data
reg_lm <- lm(Total.Claim.Amount~Number.of.Open.Complaints+EmploymentStatus+Location.Code+Gender,data=tra

summary(reg_lm)
```

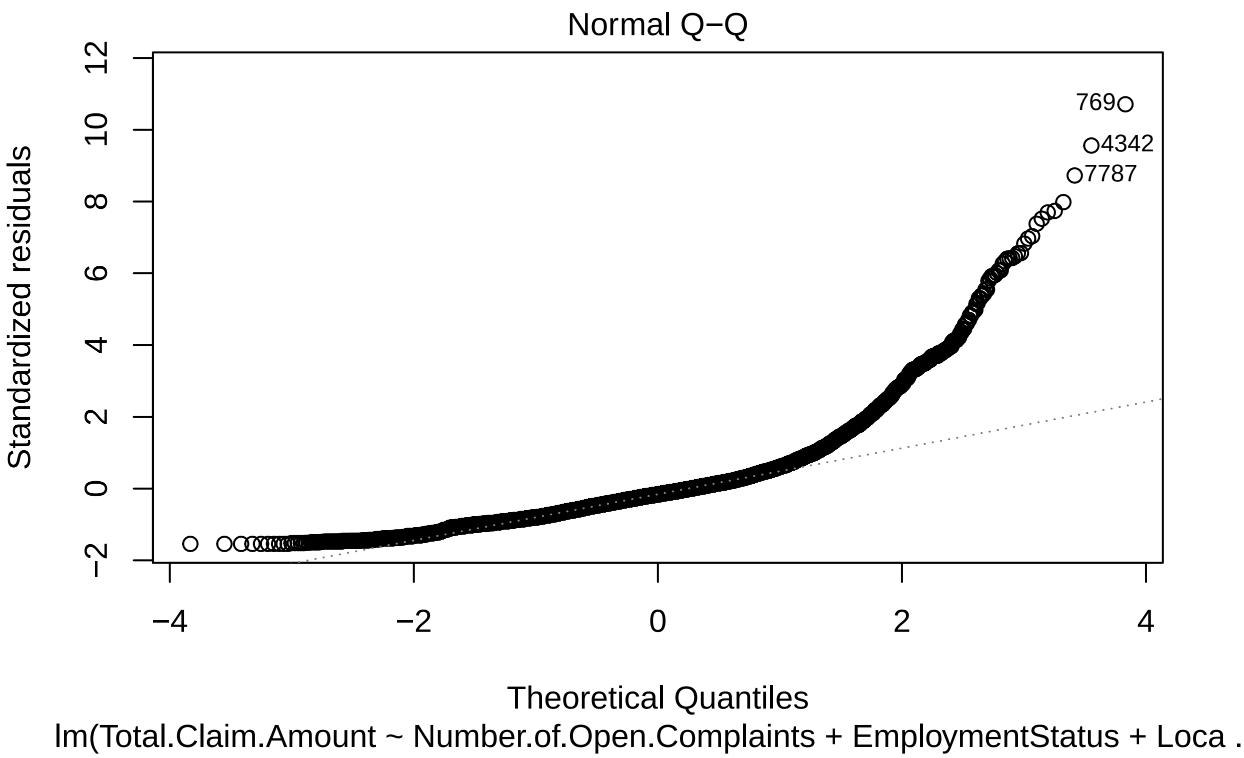
```
##
## Call:
## lm(formula = Total.Claim.Amount ~ Number.of.Open.Complaints +
##     EmploymentStatus + Location.Code + Gender, data = train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -342.50 -131.59  -35.81   60.99 2376.86
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   105.077     13.440   7.818 6.06e-15 ***
## Number.of.Open.Complaints      -6.707      2.788  -2.406  0.01615 *
## EmploymentStatusEmployed       -4.408     12.385  -0.356  0.72188
## EmploymentStatusMedical Leave   -9.454     16.543  -0.571  0.56768
## EmploymentStatusRetired       -43.438     18.426  -2.357  0.01843 *
## EmploymentStatusUnemployed      96.054     12.810   7.499 7.16e-14 ***
## Location.CodeSuburban         419.920      7.007  59.928  < 2e-16 ***
## Location.CodeUrban            220.780      8.278  26.670  < 2e-16 ***
## GenderM                        14.254      5.040   2.828  0.00469 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 222.3 on 7851 degrees of freedom
## Multiple R-squared:  0.4075, Adjusted R-squared:  0.4069
## F-statistic: 675.1 on 8 and 7851 DF,  p-value: < 2.2e-16
```

```
##build a generalized linear regression model with gamma
reg_glm <- glm(Total.Claim.Amount~Number.of.Open.Complaints+EmploymentStatus+Location.Code+Gender,data=t

summary(reg_glm)
```

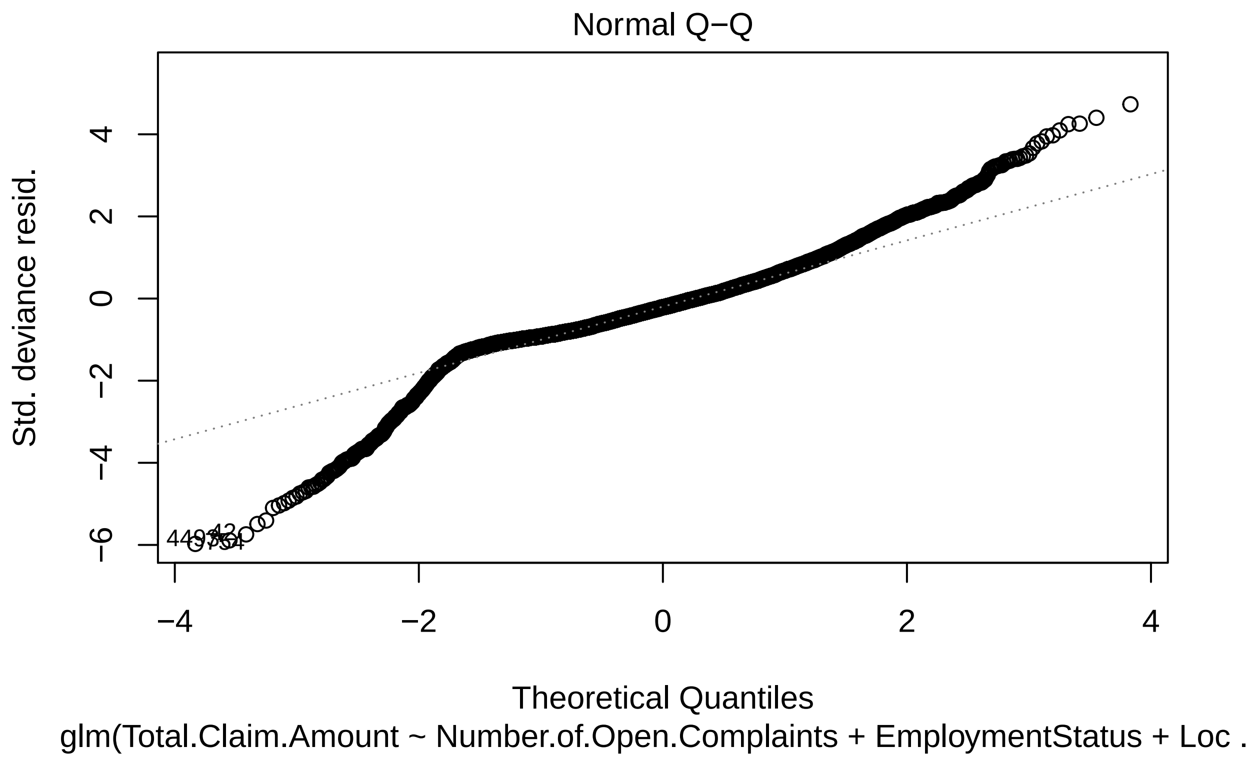
```
##
## Call:
```

```
## glm(formula = Total.Claim.Amount ~ Number.of.Open.Complaints +
##     EmploymentStatus + Location.Code + Gender, family = Gamma(link = "log"),
##     data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0598  -0.3797  -0.1087   0.1775   2.4200
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.677785   0.030981 150.989  < 2e-16 ***
## Number.of.Open.Complaints     -0.011025   0.006426  -1.716   0.0863 .
## EmploymentStatusEmployed       0.007481   0.028548   0.262   0.7933
## EmploymentStatusMedical Leave -0.020617   0.038134  -0.541   0.5888
## EmploymentStatusRetired       -0.072823   0.042474  -1.715   0.0865 .
## EmploymentStatusUnemployed     0.175889   0.029528   5.957 2.68e-09 ***
## Location.CodeSuburban          1.570744   0.016152  97.248  < 2e-16 ***
## Location.CodeUrban             1.098693   0.019082  57.577  < 2e-16 ***
## GenderM                        0.030488   0.011617   2.624   0.0087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.2625258)
##
##     Null deviance: 4398.2  on 7859  degrees of freedom
## Residual deviance: 2021.4  on 7851  degrees of freedom
## AIC: 103033
##
## Number of Fisher Scoring iterations: 5
```

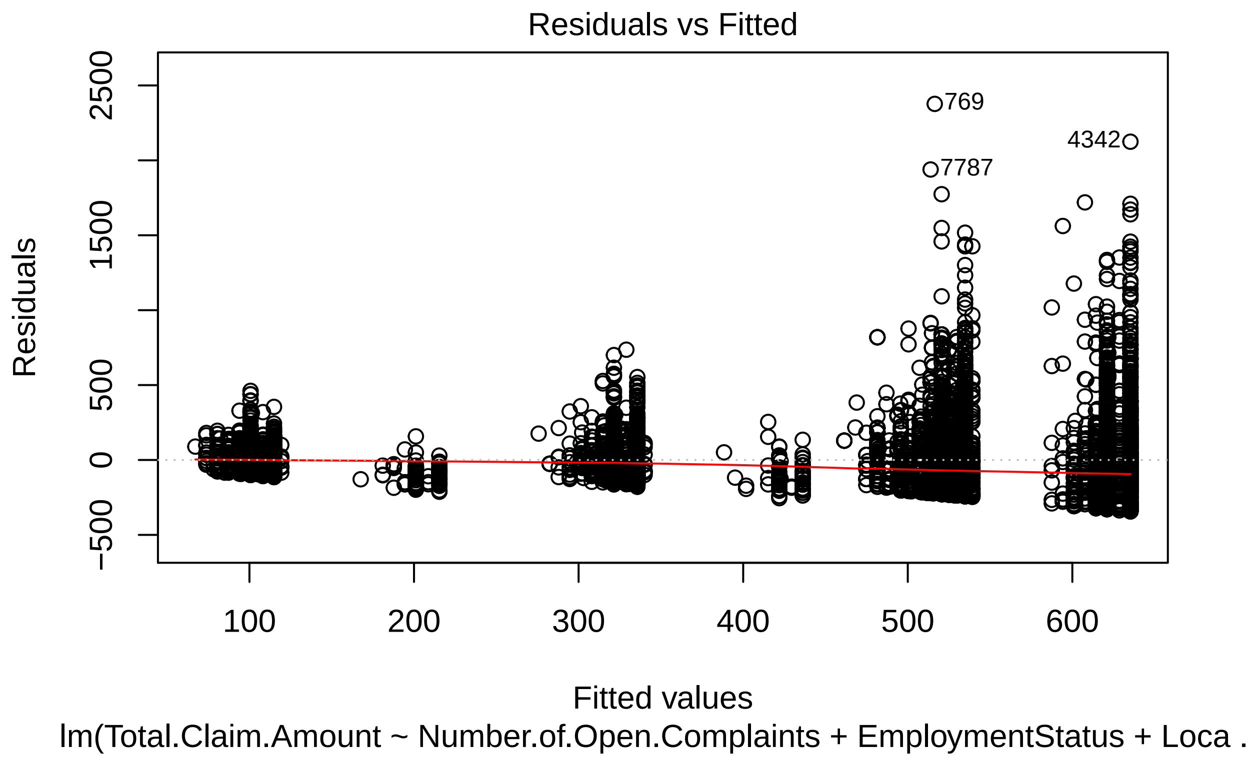
```
##Both models point out the high significant level of Employment Status, accident location, gender. Une


##test normality of deviance
plot(reg_lm,2)
```



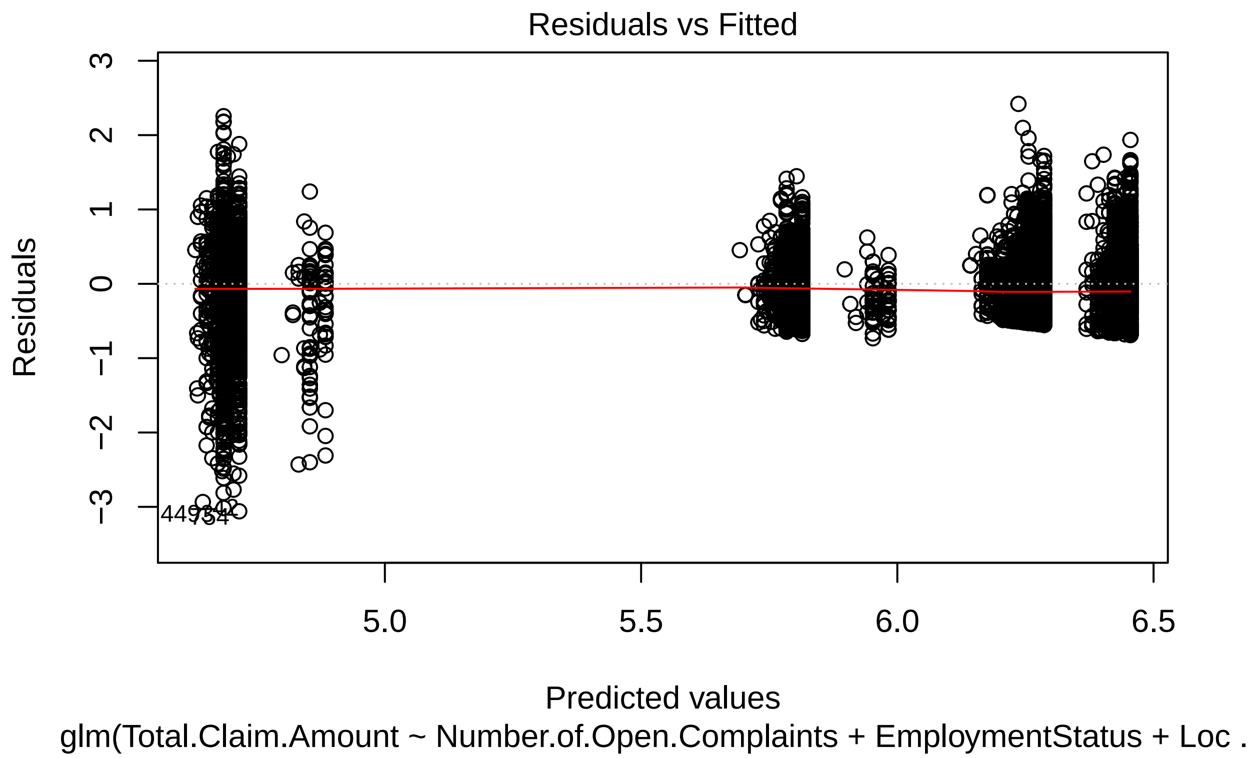

```
plot(reg_glm,2)
```



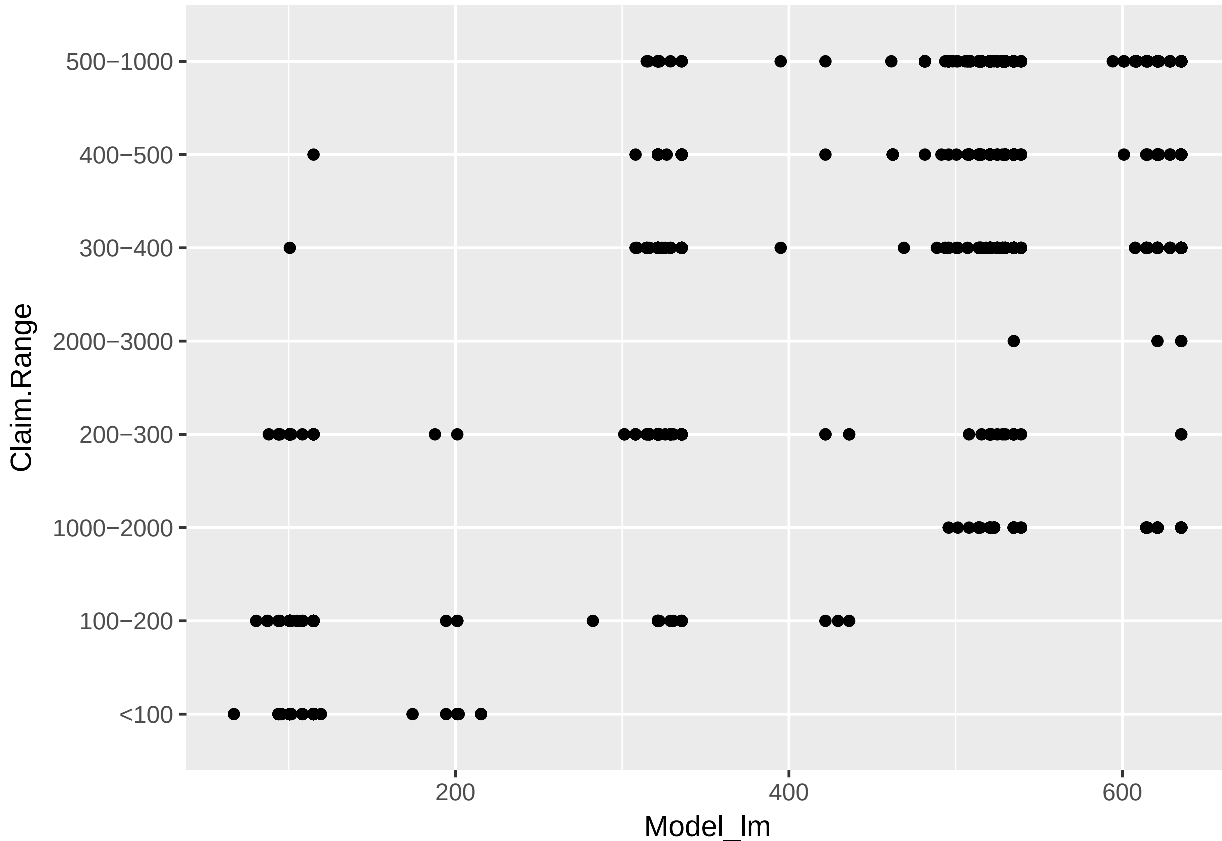

```r
plot(reg_lm,1)
```



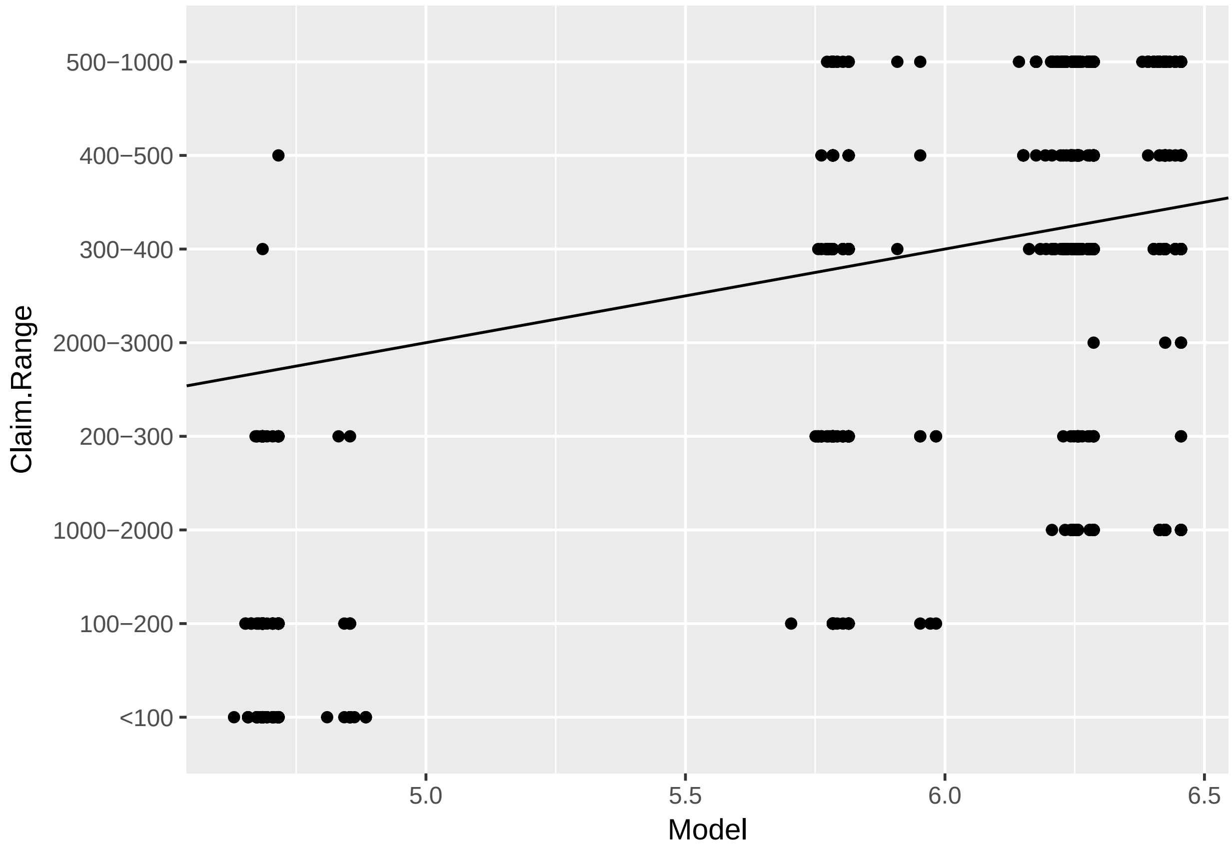

```
plot(reg_glm,1)
```



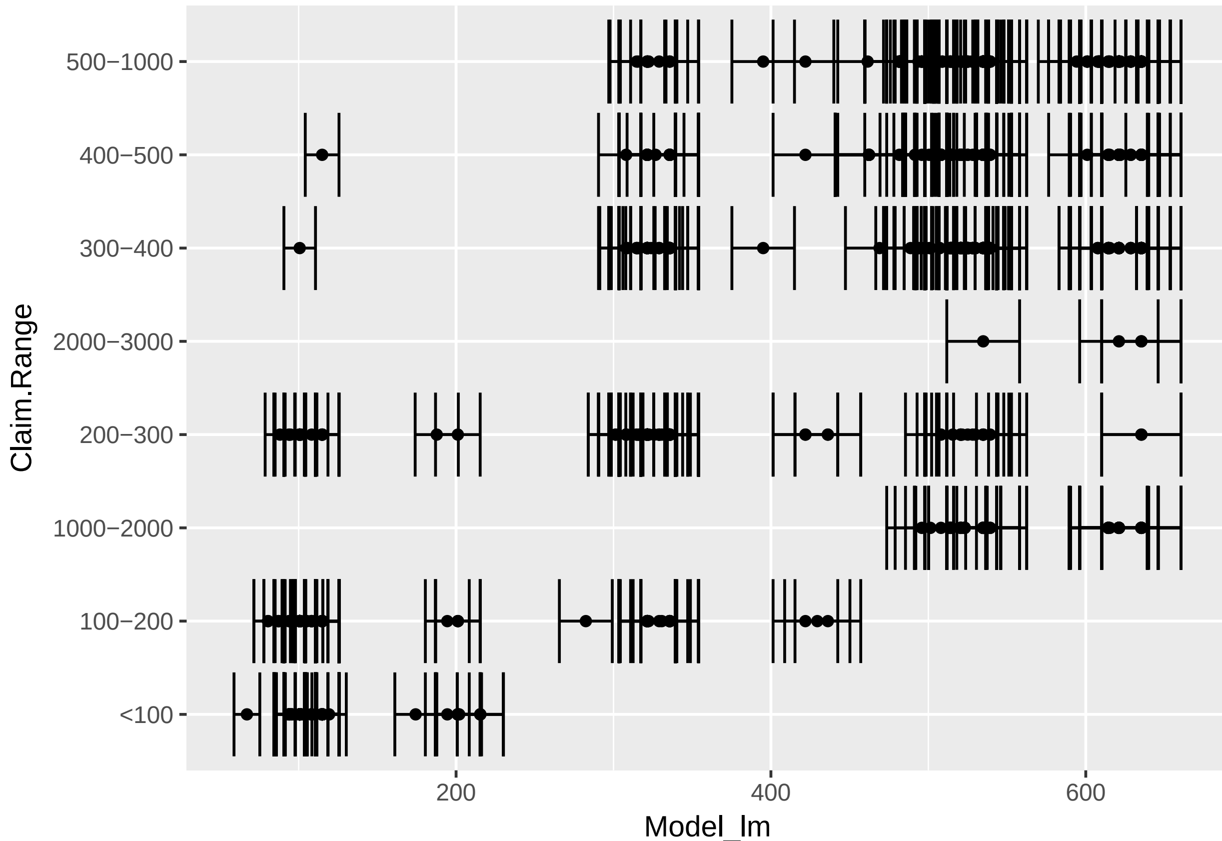

```
##raditional Linear Regression Model has less normality than Generlized Regression Model

##predict
test.data$Model_lm <- predict(reg_lm,newdata=test.data)
test.data$Model <- predict(reg_glm,newdata=test.data)

##graph
ggplot(data=test.data,aes(x=Model_lm,y=Claim.Range))+geom_point()+geom_abline(slope=1)
```



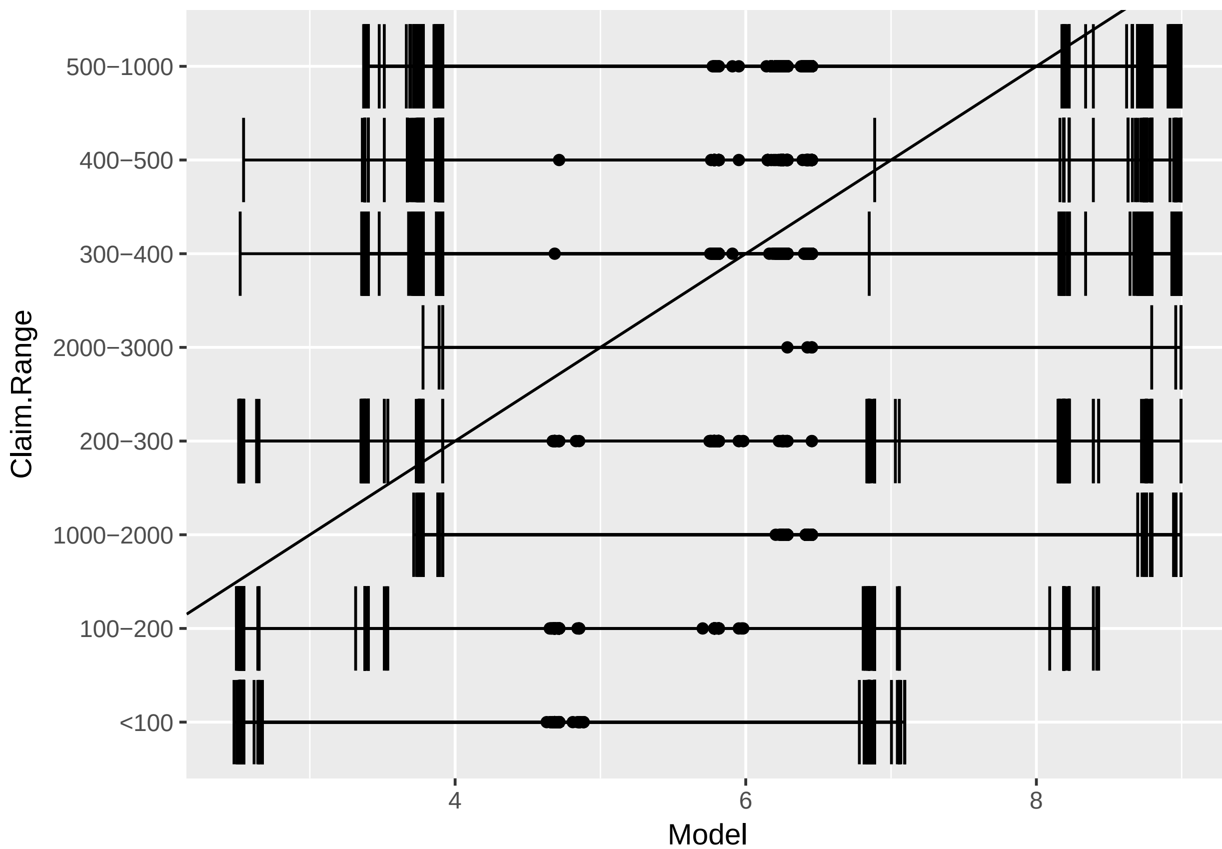

```
ggplot(data=test.data,aes(x=Model,y=Claim.Range))+geom_point()+geom_abline(slope=1)
```




```
##model error
test.data$ModelErr_lm <- sqrt(predict(reg_lm,newdata=test.data,type ='response'))

test.data$ModelErr <- sqrt(predict(reg_glm,newdata=test.data,type ='link'))

##graph
ggplot(data=test.data,aes(x=Model_lm,y=Claim.Range))+
    geom_point()+
  geom_errorbarh(aes(xmin=Model_lm-ModelErr_lm,xmax=Model_lm+ModelErr_lm))+
  geom_abline(slope=1)
```




```
ggplot(data=test.data,aes(x=Model,y=Claim.Range))+
    geom_point()+
  geom_errorbarh(aes(xmin=Model-ModelErr,xmax=Model+ModelErr))+
  geom_abline(slope=1)
```




```
##Conclusion
##Linear regression model is less accurate than Genearlized linear regression model
##The biggest challenge is to understand the statistic terms, functions, and output of statistical model

##reference
##http://actuaries.org.sg/files/library/forum_presentation/2011/2011%20Talks/11%200616%20PredictiveModel
##https://www.youtube.com/watch?v=0gf5iLTbiQM&t=10737s
##https://www.r-bloggers.com/generalized-linear-models-for-predicting-rates/
```