STAT 151A hw7

Esther Xuanpei Ouyang

4/22/2017

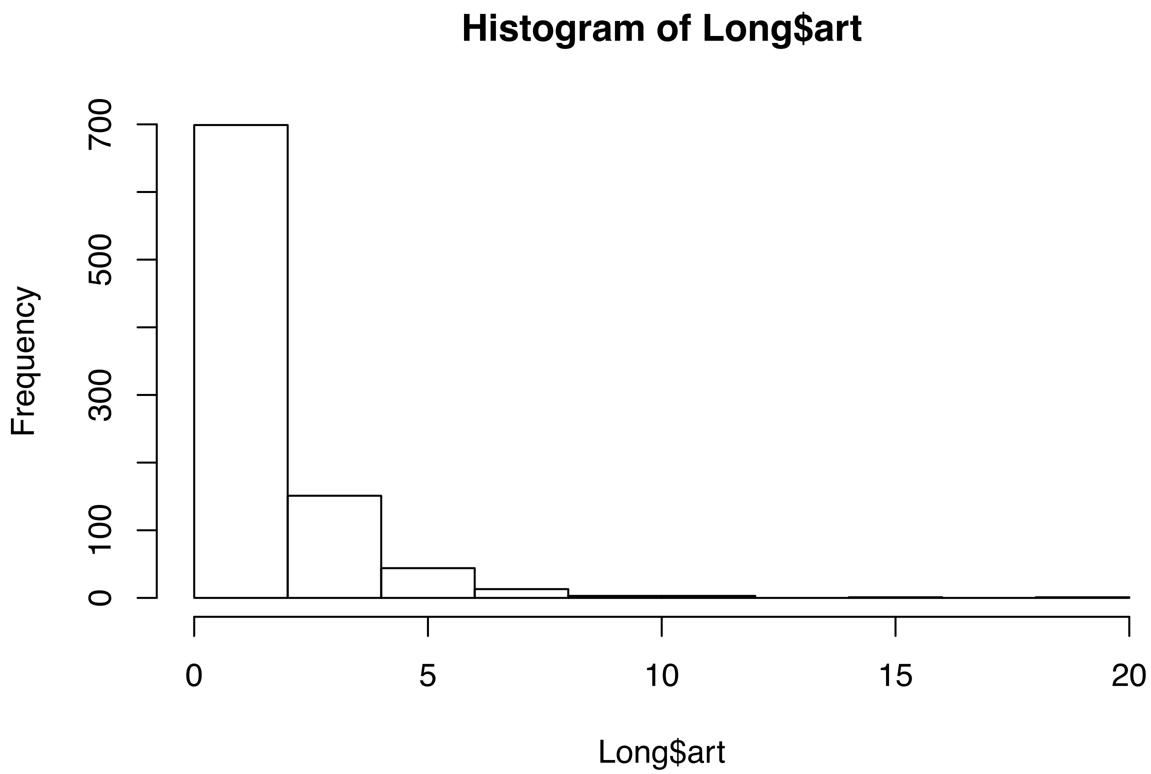
Exercise D15.1

(a) Examine the distribution of the response variable. Based on this distribution, does it appear promising to model these data by linear least-squares regression, perhaps after transforming the response? Explain your answer.

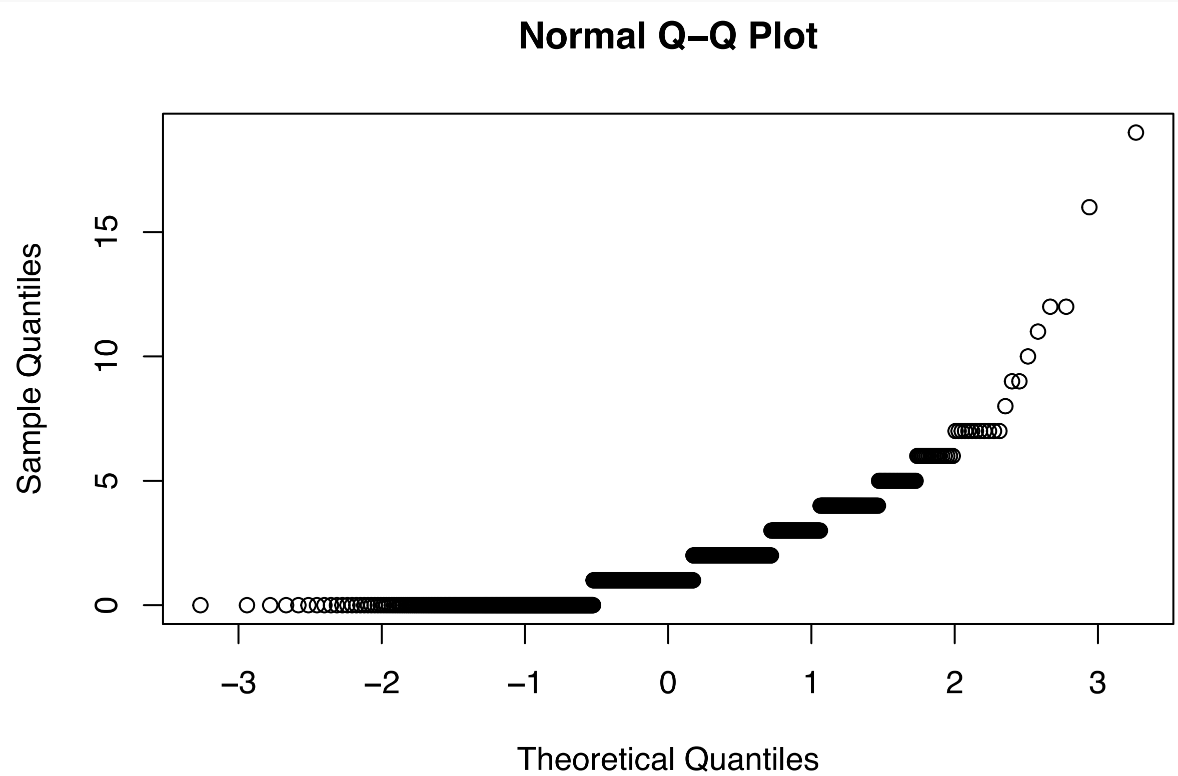
```r
Long = read.table("~/Desktop/STAT 151A/STAT-151A/hw/hw7/Long.txt")
summary(Long)
```

```
##       fem              ment             phd             mar        
##  Min.   :0.0000   Min.   : 0.000   Min.   :0.760   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.: 3.000   1st Qu.:2.260   1st Qu.:0.0000  
##  Median :0.0000   Median : 6.000   Median :3.150   Median :1.0000  
##  Mean   :0.4601   Mean   : 8.767   Mean   :3.103   Mean   :0.6623  
##  3rd Qu.:1.0000   3rd Qu.:12.000   3rd Qu.:3.920   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :77.000   Max.   :4.620   Max.   :1.0000  
##       kid5             art        
##  Min.   :0.0000   Min.   : 0.000  
##  1st Qu.:0.0000   1st Qu.: 0.000  
##  Median :0.0000   Median : 1.000  
##  Mean   :0.4951   Mean   : 1.693  
##  3rd Qu.:1.0000   3rd Qu.: 2.000  
##  Max.   :3.0000   Max.   :19.000
```

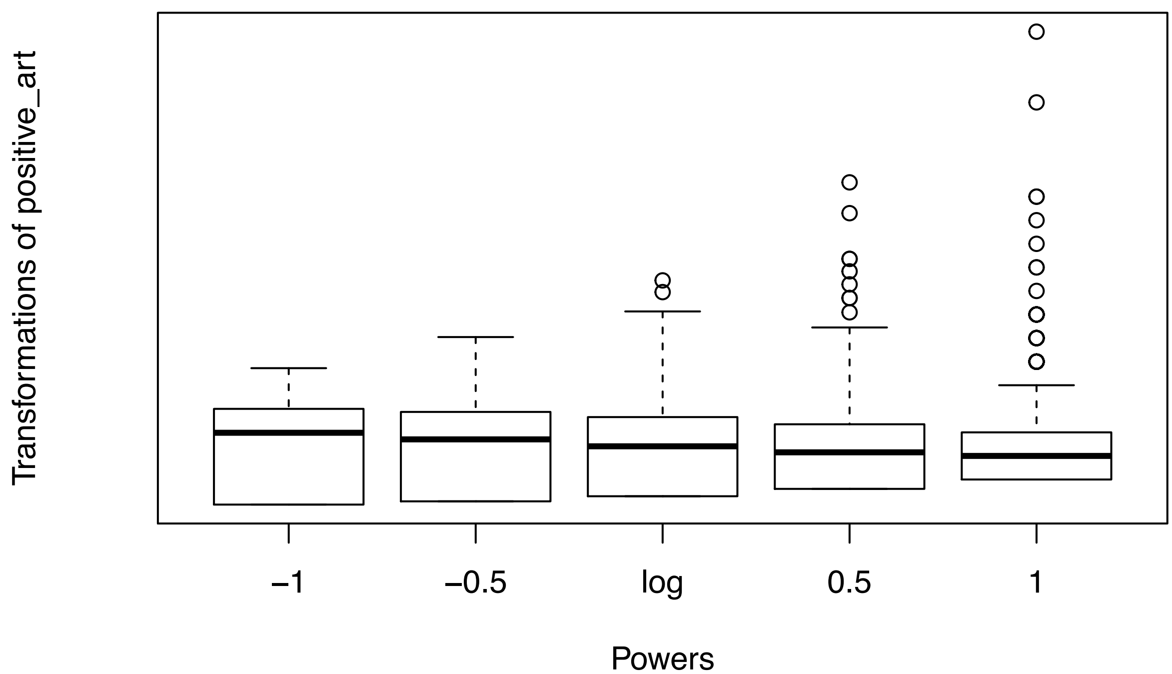
```r
attach(Long)

# Plot the original response variable
hist(Long$art)
```



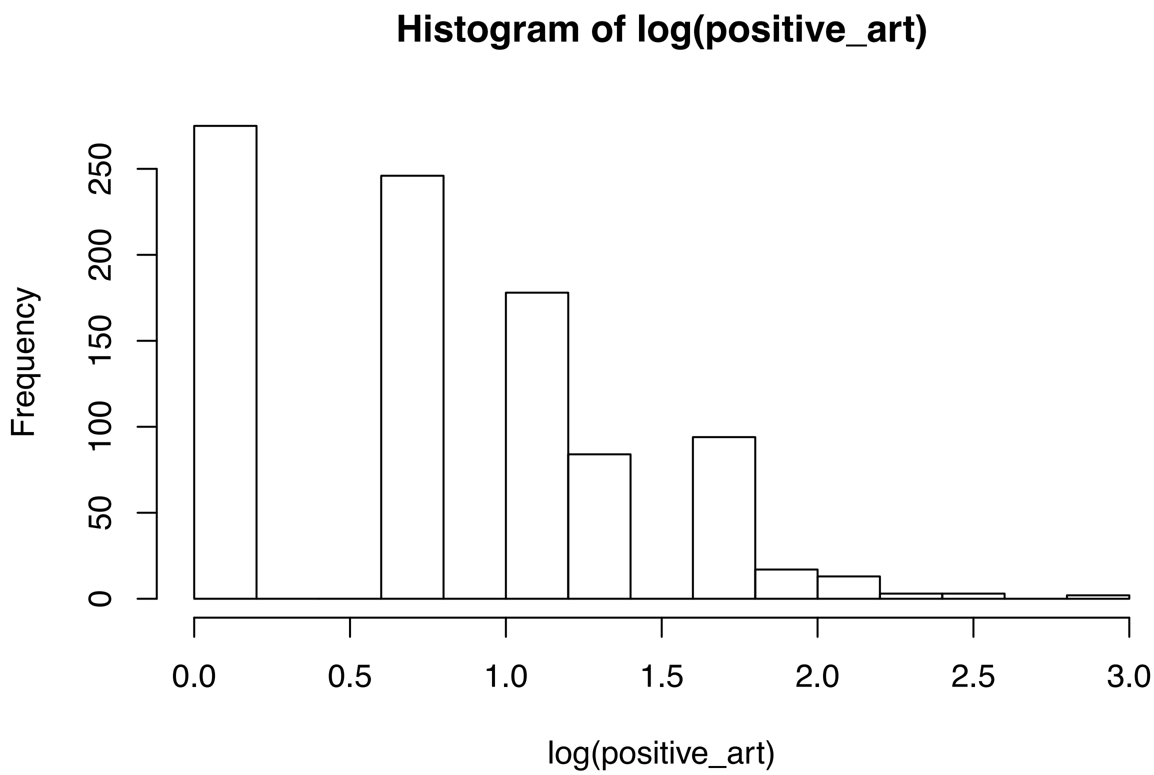

```
qqnorm(Long$art)
```



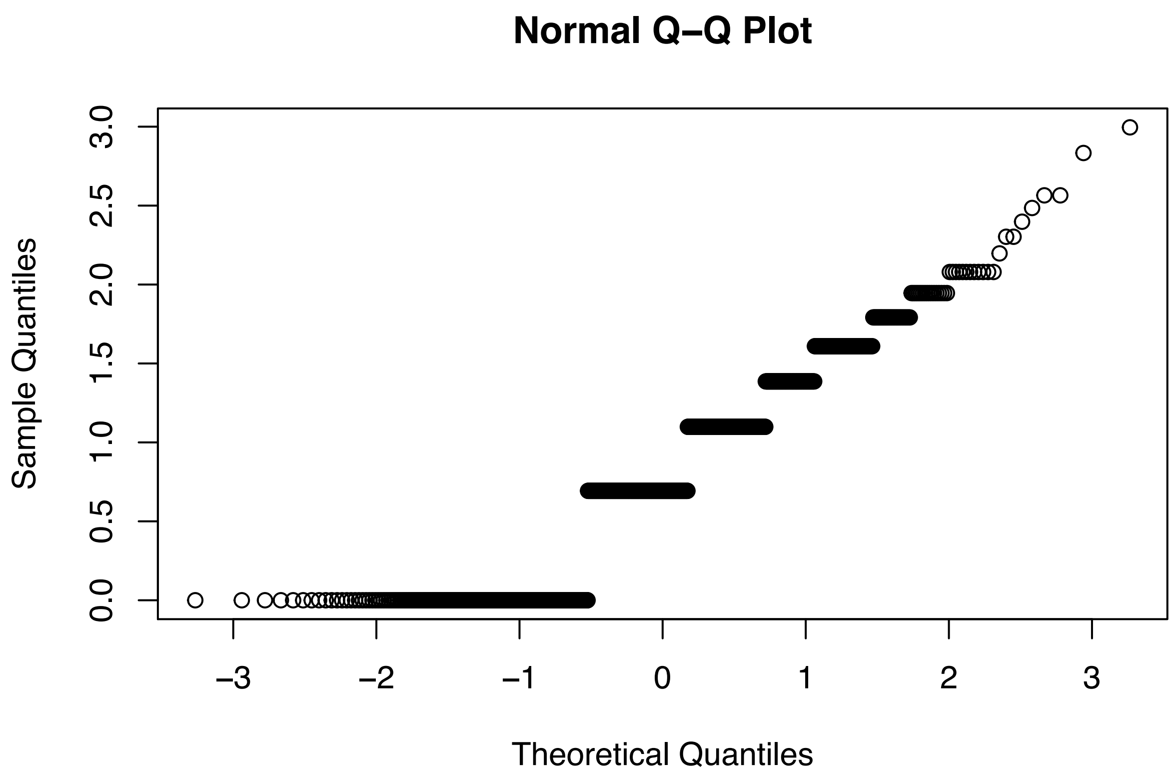

```
positive_art = Long$art + 1
symbox(~positive_art, data = Long)
```



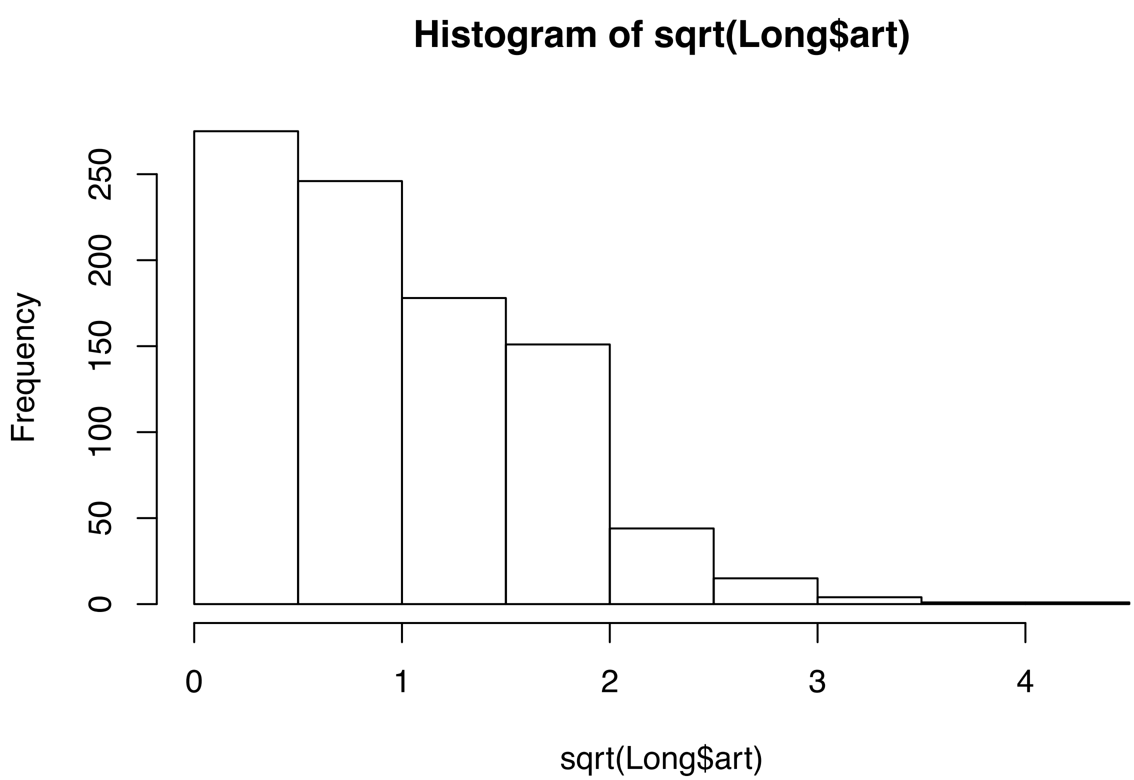

```r
# Plot the log transformed response variable
hist(log(positive_art))
```



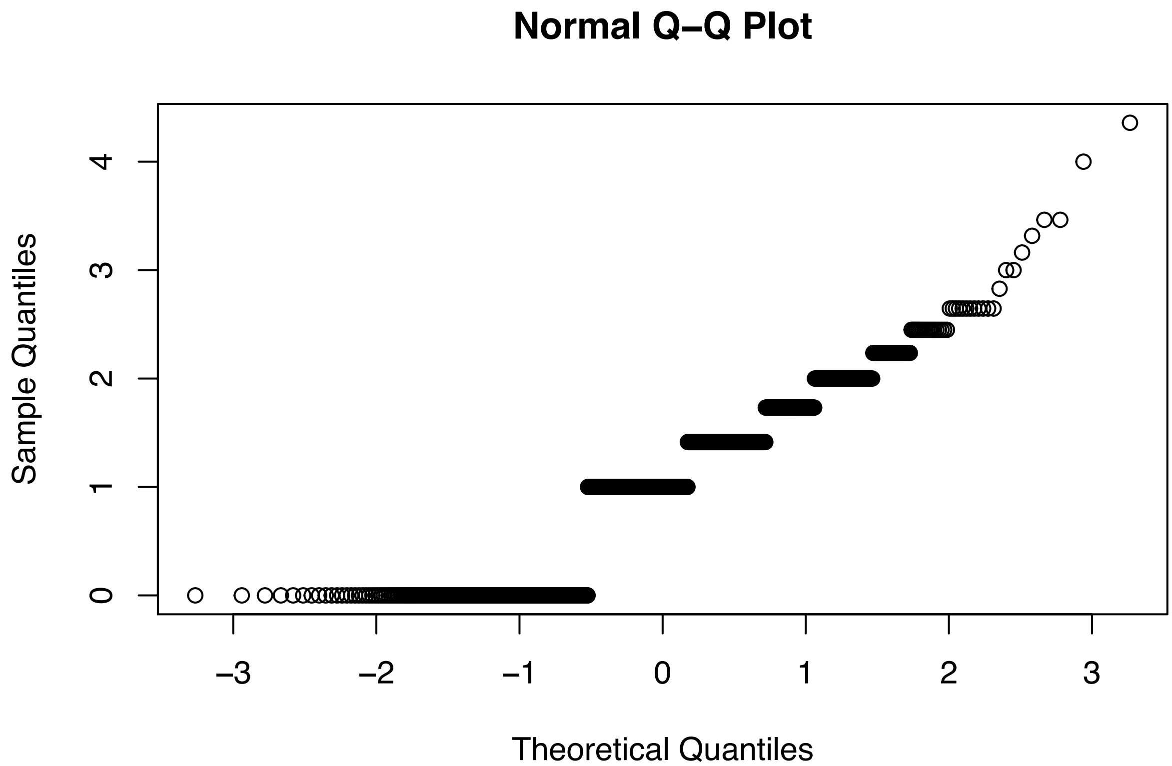

```r
qqnorm(log(positive_art))
```




```r
# Plot the log transformed response variable
hist(sqrt(Long$art))
```



```r
qqnorm(sqrt(Long$art))
```



Based on the histogram and the normal probability plot of the response variable, we can find out the data is not normally distributed but very right skewed. Instead, the distribution of response variable looks like a poisson distribution with small lambda. Also, the data looks more normal if we log or square root transform it.

Therefore, the result will not looks promising if we fit the data by linear least-sqaure regression.

(b) Following Long, perform a Poisson regression of art on the explanatory variables. What do you conclude from the results of this regression?

```r
fit = glm(art ~ fem+ment+phd+mar+kid5+art, family = poisson(), data = Long)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 6 in
## model.matrix: no columns are assigned
```

```r
summary(fit) # display results
```

```
##
## Call:
## glm(formula = art ~ fem + ment + phd + mar + kid5 + art, family = poisson(),
##     data = Long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5672  -1.5398  -0.3660   0.5722   5.4467
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.304619   0.102980   2.958   0.0031 **
## fem         -0.224594   0.054613  -4.112 3.92e-05 ***
## ment         0.025543   0.002006  12.733  < 2e-16 ***
## phd          0.012822   0.026397   0.486   0.6271
## mar          0.155243   0.061374   2.529   0.0114 *
## kid5        -0.184883   0.040127  -4.607 4.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
## AIC: 3314.1
##
## Number of Fisher Scoring iterations: 5
```

```r
confint(fit) # 95% CI for the coefficients
```

```
## Waiting for profiling to be done...

## Warning in model.matrix.default(fitted, data = structure(list(art = c(3L, :
## the response appeared on the right-hand side and was dropped

## Warning in model.matrix.default(fitted, data = structure(list(art = c(3L, :
## problem with term 6 in model.matrix: no columns are assigned

##                   2.5 %      97.5 %
## (Intercept)  0.10156004  0.50526499
## fem         -0.33193992 -0.11781816
## ment         0.02154163  0.02940716
## phd         -0.03881094  0.06467394
## mar          0.03520217  0.27584820
## kid5        -0.26422745 -0.10689873
```

```r
exp(coef(fit)) # exponentiated coefficients
```

```
## (Intercept)         fem        ment         phd         mar        kid5
##   1.3561083   0.7988403   1.0258718   1.0129045   1.1679420   0.8312018
```

```r
exp(confint(fit)) # 95% CI for exponentiated coefficients
```

```
## Waiting for profiling to be done...

## Warning in model.matrix.default(fitted, data = structure(list(art = c(3L, :
## the response appeared on the right-hand side and was dropped

## Warning in model.matrix.default(fitted, data = structure(list(art = c(3L, :
## problem with term 6 in model.matrix: no columns are assigned

##                 2.5 %    97.5 %
## (Intercept) 1.1068964 1.6574247
## fem         0.7175304 0.8888577
## ment        1.0217753 1.0298438
## phd         0.9619326 1.0668111
## mar         1.0358291 1.3176478
```

```
## kid5        0.7677989 0.8986167
```

```
fit_predicted = predict(fit, type="response") # predicted values
fit_residuals = residuals(fit, type="deviance") # residuals
```

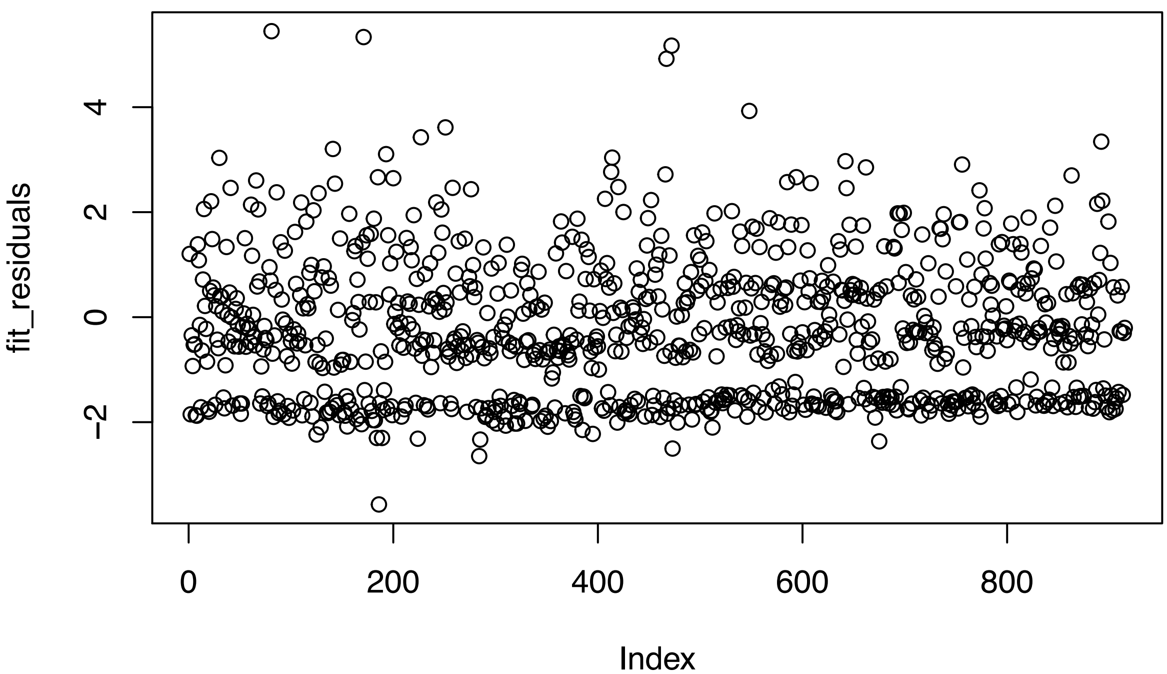
In terms of hypothesis testing with significance level at 0.05, the explanatory variables fem, ment, kid5 are significant while phd and mar are insignificant. The range of deviance residual is not large. However, we notice that there is overdispersion because the residual deviance is much greater than the degrees of freedom.

(c) Perform regression diagnostics on the model fit in the previous question. If you identify any problems, try to deal with them. Are the conclusions of the research altered?

```
# Assessing Outiers
outlierTest(fit)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 81  5.513096         3.5258e-08   3.2261e-05
## 171 5.361646         8.2467e-08   7.5457e-05
## 472 5.208086         1.9080e-07   1.7458e-04
## 467 5.085126         3.6738e-07   3.3615e-04
```

```
# analysis variance of error
plot(fit_residuals)
```




```
# variance inflation factors
vif(fit)
```

```
## Warning in model.matrix.default(mod, data = structure(list(art = c(3L,
## 0L, : the response appeared on the right-hand side and was dropped

## Warning in model.matrix.default(mod, data = structure(list(art = c(3L,
## 0L, : problem with term 6 in model.matrix: no columns are assigned

##          GVIF Df GVIF^(1/(2*Df))
## fem  1.108477  1        1.052842
```

```
## ment 1.081111  1        1.039765
## phd  1.067309  1        1.033107
## mar  1.264643  1        1.124563
## kid5 1.286111  1        1.134068
## art  1.411834  0             Inf
```

The variance inflation factors is not large for explanatory variables fem, ment, phd, mar and kid5 but infinity for response variable art, which means the explanatory variables are not correlated and the response variables is highly correlated with the other explanantory variables.

By plotting out the residual, we can see the variance of error is approximately constant.

Also, in the model fit in (b), since the residual deviance is much larger than the degree of freedom, we can overdispersion and need to use quasipoisson() instead of poisson().

(d) Refit Long's model allowing for overdispersion (using a quasi-Poisson or negative-binomial model). Does this make a difference to the results?

```r
overdisper_fit = glm(art ~ fem+ment+phd+mar+kid5+art, family = quasipoisson(link = "log"), data = Long)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 6 in
## model.matrix: no columns are assigned
```

```r
summary(overdisper_fit) # display results
```

```
##
## Call:
## glm(formula = art ~ fem + ment + phd + mar + kid5 + art, family = quasipoisson(link = "log"),
##     data = Long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5672  -1.5398  -0.3660   0.5722   5.4467
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.304619   0.139271   2.187 0.028979 *
## fem         -0.224594   0.073860  -3.041 0.002427 **
## ment         0.025543   0.002713   9.415  < 2e-16 ***
## phd          0.012822   0.035699   0.359 0.719552
## mar          0.155243   0.083003   1.870 0.061759 .
## kid5        -0.184883   0.054268  -3.407 0.000686 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.829006)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
confint(overdisper_fit) # 95% CI for the coefficients
```

```
## Waiting for profiling to be done...

## Warning in model.matrix.default(fitted, data = structure(list(art = c(3L, :
## the response appeared on the right-hand side and was dropped

## Warning in model.matrix.default(fitted, data = structure(list(art = c(3L, :
## problem with term 6 in model.matrix: no columns are assigned

##                    2.5 %      97.5 %
## (Intercept)  0.029409303  0.57541492
## fem         -0.369927610 -0.08030316
## ment         0.020098040  0.03073753
## phd         -0.056961086  0.08300571
## mar         -0.007001095  0.31851958
## kid5        -0.292532116 -0.07972242
```

```r
exp(coef(overdisper_fit)) # exponentiated coefficients
```

```
## (Intercept)         fem        ment         phd         mar        kid5
##   1.3561083   0.7988403   1.0258718   1.0129045   1.1679420   0.8312018
```

```r
exp(confint(overdisper_fit)) # 95% CI for exponentiated coefficients
```

```
## Waiting for profiling to be done...

## Warning in model.matrix.default(fitted, data = structure(list(art = c(3L, :
## the response appeared on the right-hand side and was dropped

## Warning in model.matrix.default(fitted, data = structure(list(art = c(3L, :
## problem with term 6 in model.matrix: no columns are assigned

##                 2.5 %    97.5 %
## (Intercept) 1.0298460 1.7778681
## fem         0.6907843 0.9228365
## ment        1.0203014 1.0312148
## phd         0.9446308 1.0865480
## mar         0.9930234 1.3750905
## kid5        0.7463713 0.9233726
```

```r
predicted = predict(overdisper_fit, type="response") # predicted values
quasifit_residual = residuals(overdisper_fit, type="deviance") # residuals

overdisper_fit_sqrt = glm(art ~ fem+ment+phd+mar+kid5+art, family = quasipoisson(link = "sqrt"), data =
```

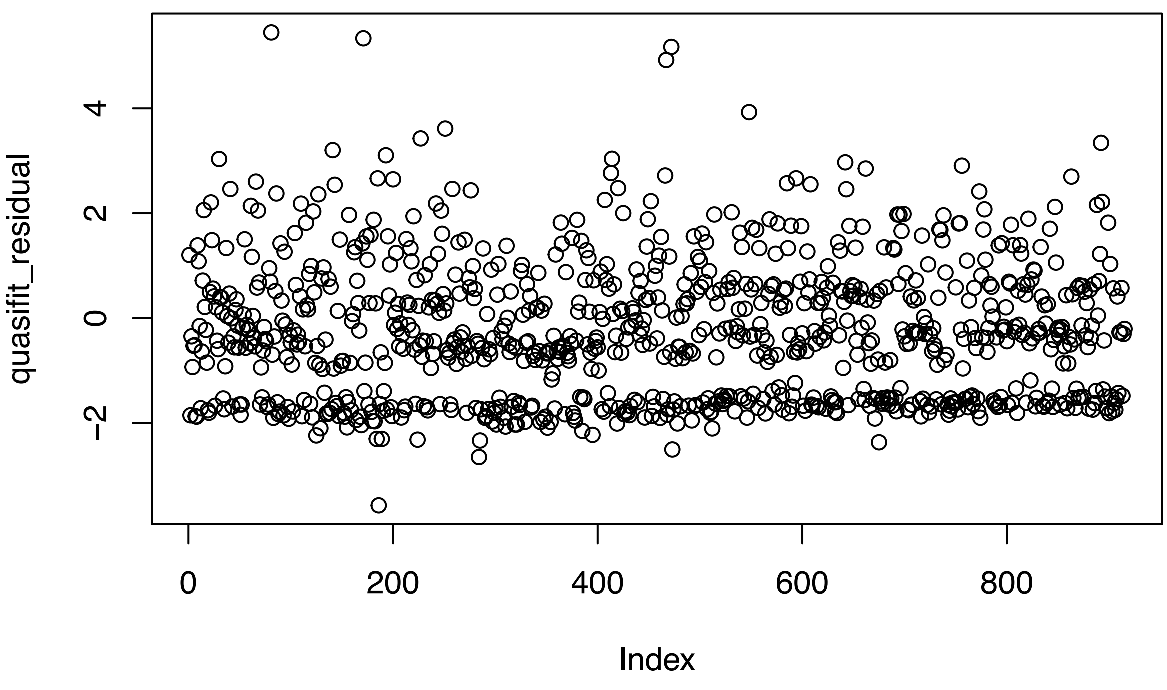
```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 6 in
## model.matrix: no columns are assigned
```

```r
summary(overdisper_fit_sqrt)
```

```
##
## Call:
## glm(formula = art ~ fem + ment + phd + mar + kid5 + art, family = quasipoisson(link = "sqrt"),
##     data = Long)
##
## Deviance Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -3.0809  -1.4929  -0.3535   0.5979   5.3701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.126411   0.089226  12.624  < 2e-16 ***
## fem         -0.138636   0.047460  -2.921 0.003574 **
## ment         0.021674   0.002446   8.859  < 2e-16 ***
## phd          0.005702   0.023580   0.242 0.808962
## mar          0.099000   0.053862   1.838 0.066382 .
## kid5        -0.111546   0.033675  -3.312 0.000961 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.820918)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1620.0  on 909  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
plot(quasifit_residual)
```



Here, if we use quasi-Poisson with log link function to fit the generalized linear model to account for the overdispersion, we can see that the residual deviance does not change at all.

But if we fit the model with quasi-Poisson with sqrt link function, we can see that the residual deviance change from 1634.4 to 1620, which do gives a slightly better fitted model.

Exercise D15.2

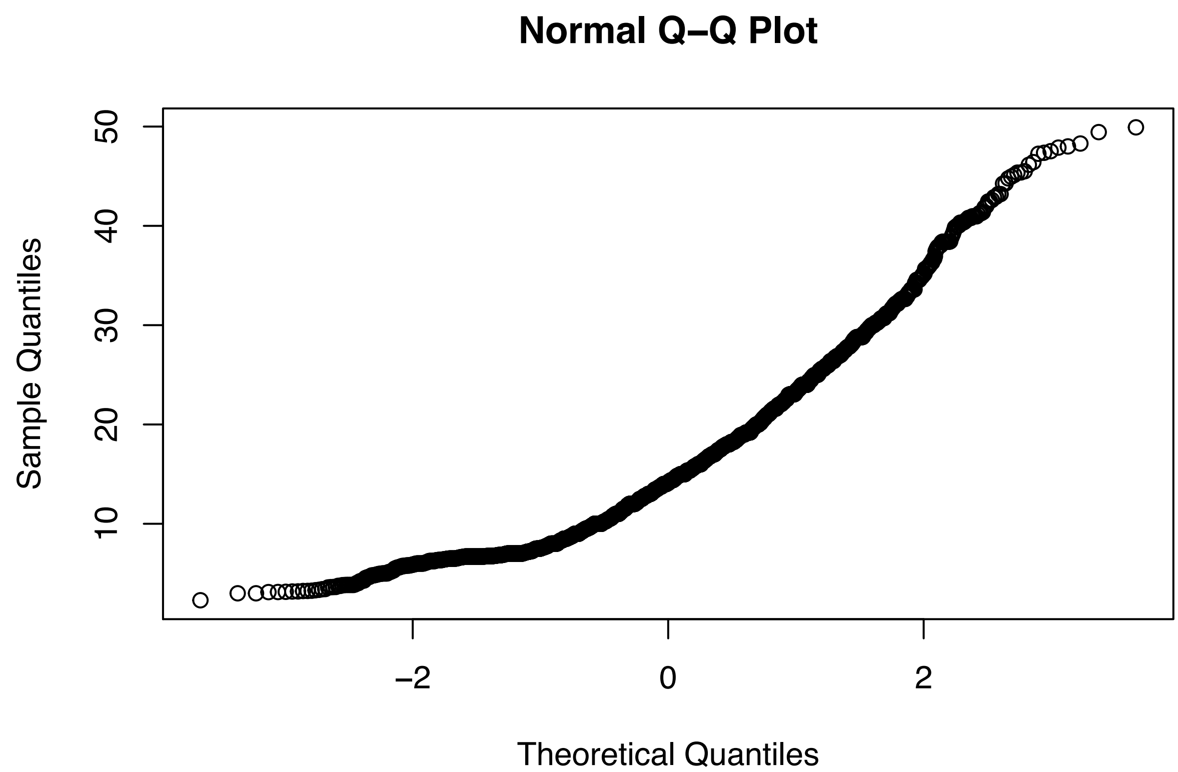
Chapter 12 describes the linear regression of wages on gender, age, and education for data drawn from the Canadian Survey of Labour and Income Dynamics (the "SLID"). The data are in the file SLID-Ontario.txt. In the text, the response variable is log-transformed to correct skewness and non-constant spread in the regression. Consider an alternative strategy employing a gamma generalized linear model. After fitting this model and checking its adequacy, which of the two approaches to the data do you prefer?

```
SLID = read.table("~/Desktop/STAT 151A/STAT-151A/hw/hw7/SLID-Ontario.txt", header = TRUE)
summary(SLID)
```

```
##       age             sex        compositeHourlyWages yearsEducation
##  Min.   :16.00   Female:2007   Min.   : 2.30        Min.   : 0.00
##  1st Qu.:28.00   Male  :1990   1st Qu.: 9.25        1st Qu.:12.00
##  Median :36.00                 Median :14.13        Median :13.00
##  Mean   :36.96                 Mean   :15.54        Mean   :13.21
##  3rd Qu.:46.00                 3rd Qu.:19.75        3rd Qu.:15.00
##  Max.   :65.00                 Max.   :49.92        Max.   :20.00
```

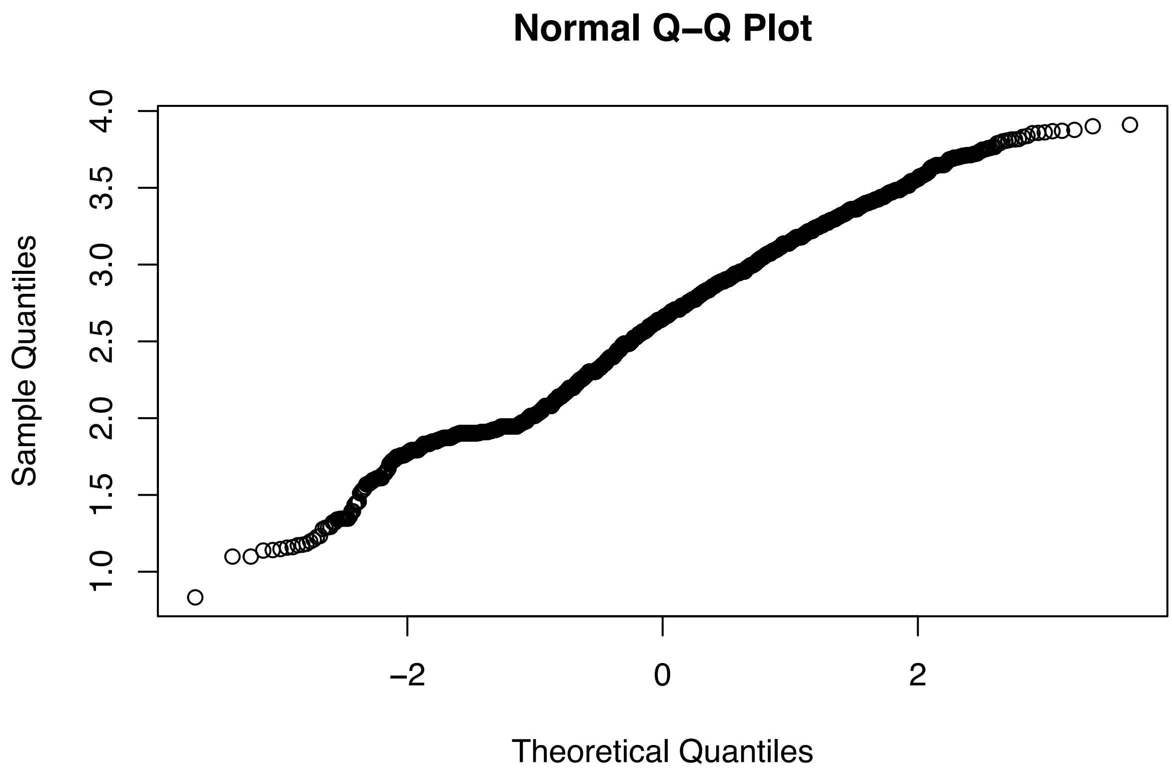
```
hist(SLID$compositeHourlyWages)
```


```
qqnorm(SLID$compositeHourlyWages)
```




```r
log_wages = log(SLID$compositeHourlyWages)
hist(log_wages)
```

```r
qqnorm(log_wages)
```



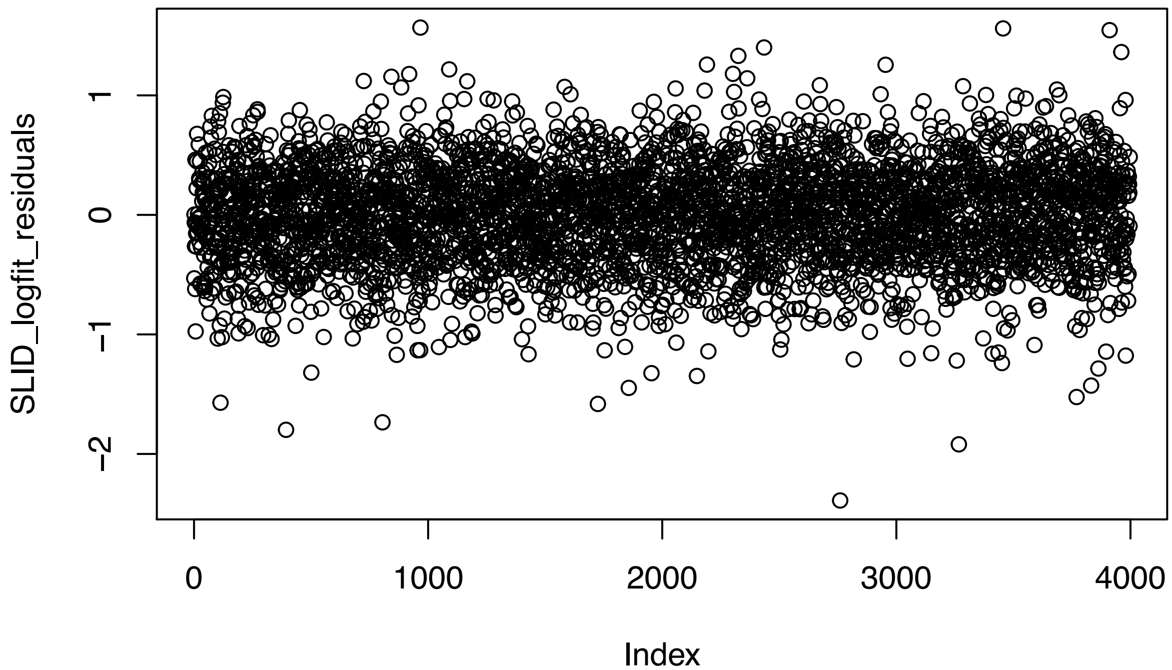

```r
fit_SLID_log = lm(log_wages~age+sex+yearsEducation, data = SLID)
summary(fit_SLID_log)
```

```
##
## Call:
## lm(formula = log_wages ~ age + sex + yearsEducation, data = SLID)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38930 -0.27670  0.01312  0.28413  1.56696
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.0990176  0.0379649   28.95   <2e-16 ***
## age            0.0181548  0.0005491   33.06   <2e-16 ***
## sexMale        0.2244959  0.0131208   17.11   <2e-16 ***
## yearsEducation 0.0558764  0.0021713   25.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4146 on 3993 degrees of freedom
## Multiple R-squared:  0.3212, Adjusted R-squared:  0.3207
## F-statistic: 629.7 on 3 and 3993 DF,  p-value: < 2.2e-16
```

```r
vif(fit_SLID_log)
```

```
##            age            sex yearsEducation
##       1.010056       1.000644       1.010440
```

```r
SLID_logfit_residuals = residuals(fit_SLID_log)
plot(SLID_logfit_residuals)
```



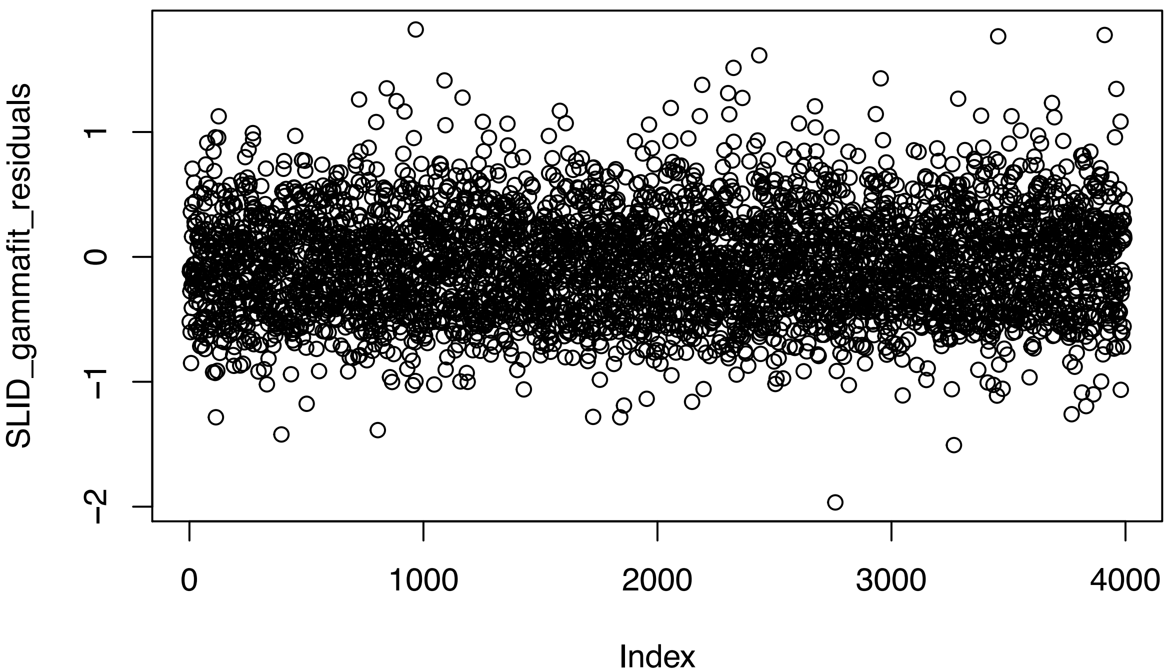

```r
fit_SLID_gamma = glm(compositeHourlyWages~age+sex+yearsEducation, family = Gamma, data = SLID)
summary(fit_SLID_gamma)
```

```
##
## Call:
## glm(formula = compositeHourlyWages ~ age + sex + yearsEducation,
##     family = Gamma, data = SLID)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9659  -0.3700  -0.0694   0.2124   1.8208
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.536e-01  2.269e-03   67.71   <2e-16 ***
## age             -9.818e-04  3.418e-05  -28.73   <2e-16 ***
## sexMale         -1.292e-02  8.647e-04  -14.95   <2e-16 ***
## yearsEducation -3.189e-03  1.241e-04  -25.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1839024)
##
##     Null deviance: 989.01  on 3996  degrees of freedom
## Residual deviance: 696.54  on 3993  degrees of freedom
## AIC: 25430
##
## Number of Fisher Scoring iterations: 5
```

```r
vif(fit_SLID_gamma)
```

```
##           age           sex yearsEducation
##      1.010334      1.015322       1.011083
```

```r
SLID_gammafit_residuals = residuals(fit_SLID_gamma)
plot(SLID_gammafit_residuals)
```



For both fitted model, all the explanatory variables are significant. Also, we can reject null hypotheses for both models. (since the F-statistics for log-transformed model is significant and for glm, residual deviance is less than null deviance).

They also have similar vif factors for their explanatory variables and similar spread of error for their residual variance.

Although the result from two models are pretty similar, I think the generalized linear regression is better since it is flexible to adjust the link function to fit the model better.