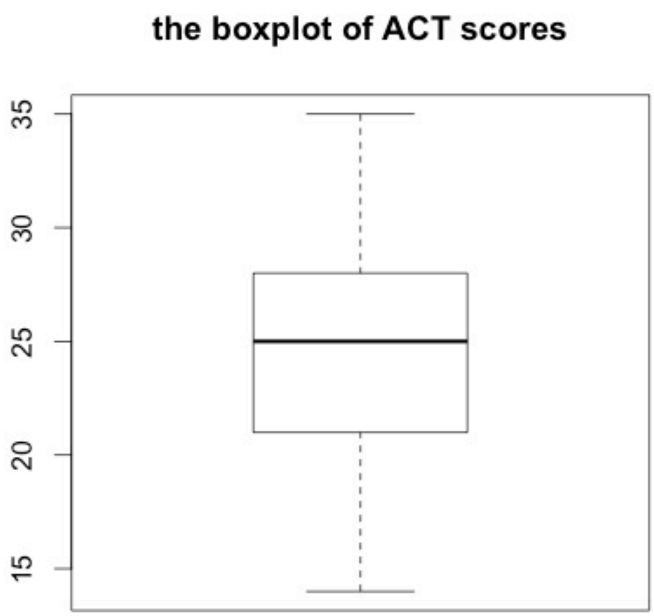
Jiahong Hu
Jh3561
HW4
STAT 4315

3.3

Part a

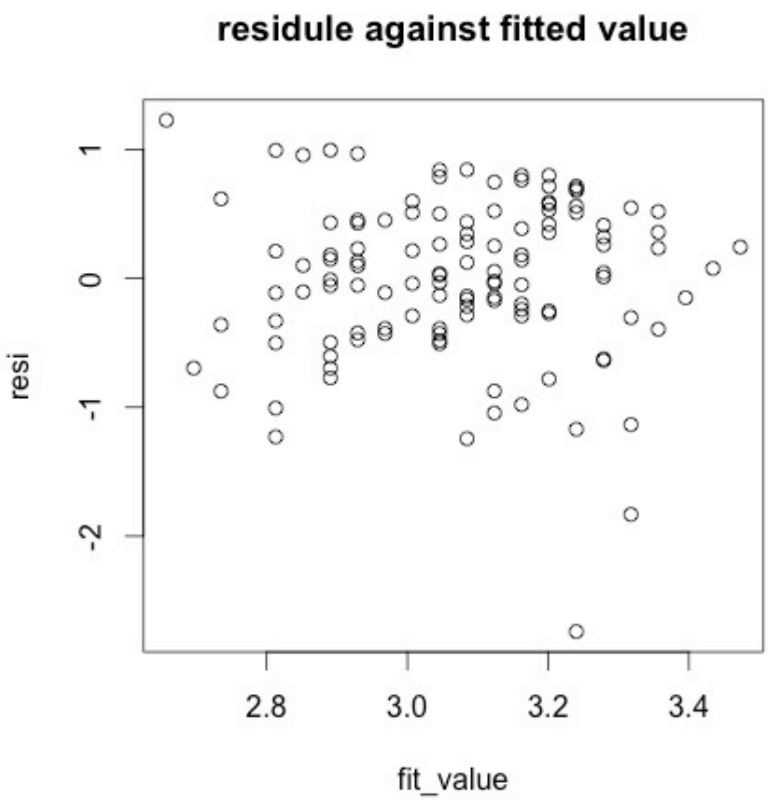
```
# 3.3
# part a |
setwd("/Users/jiahongHu/Desktop/Spring 2015/Linear Regression 4315/hw/hw4")
data<-read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdata
names(data)<-c("GPA","ACT")
attach(data)
fit<-lm(GPA~ACT)
boxplot(ACT,main="the boxplot of ACT scores")
```



The box plot shows that the ACT scores have a mean around 25, its distribution is approximately symmetric, and there are not any outliers.

Part c

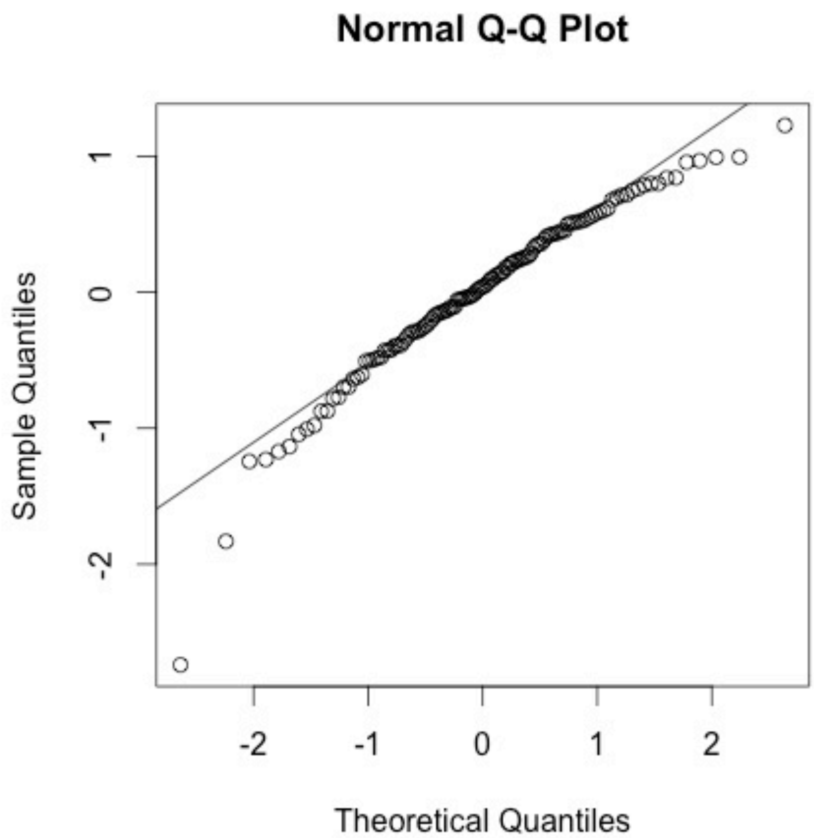
```
resi<-fit$residuals
fit_value<-fit$fitted.values
plot(fit_value,resi,main="residule against fitted value")
```



The plot shows variance of residual appears constant and does not depend on the fitted value of GPA. Most of the variances of residuals are between -2 and 1, which indicates no obvious outlier. Therefore, I think it satisfies the linearity assumption of the linear regression model and also the assumption that the error term is independent .

Part d

```
# part d
qqplot<-qqnorm(resi)
qqline(resi)
SSE = sum(resi^2)
MSE = SSE/(length(GPA)-2)
n=120
ExpVals = sapply(1:n, function(k) sqrt(MSE) * qnorm((k-.375)/(n+.25)))
cor(ExpVals,sort(fit$residuals))
```



In order to test the normality of the error distribution, I set up the hypothesis test as follows: H0: Normal, Ha: not normal.

I calculated the correlation between the ordered residuals and their expected values under the normality test. We know the correlation is r = 0.97373.

Using the table B.6 and alpha = 0.05, If r >=0.987 conclude H0, otherwise Ha.

In this case, 0.9737<0.987, I conclude Ha. There is some departure from normality.

Part e

H0: the error variance is constant (error variance is independent with X)
Ha: the error variance is not constant (error variance changes with the level of X)

Alpha = 0.01
The decision rule is: if $|t^*_{BF}| \leq t(1 - alpha/2; n - 2) = t(0.995; 118) = 2.61814$, conclude H0. Otherwise, conclude Ha

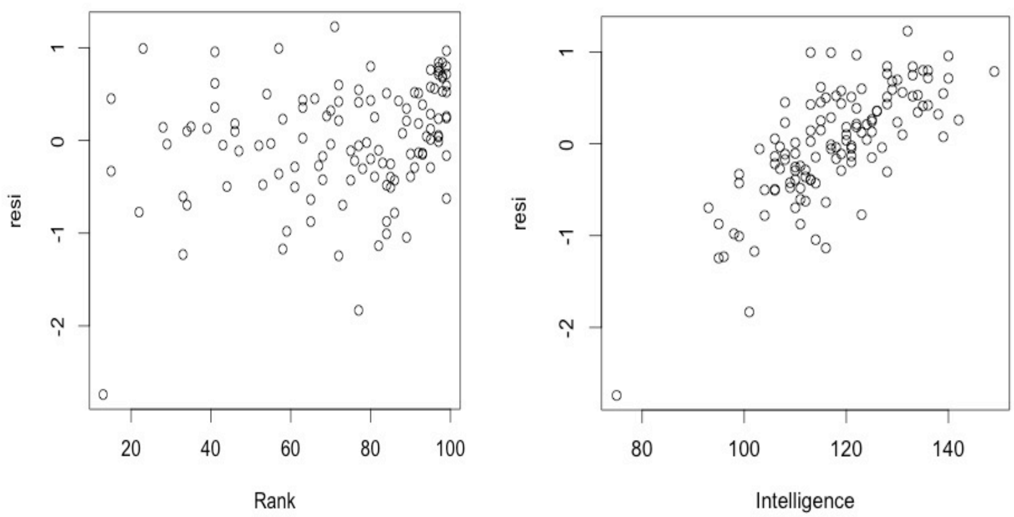
Code:

```
> t_bf
[1] -0.8967448

# part e
ACT_1<-ACT[ACT<26]
ACT_2<-ACT[ACT>=26]
n1<-length(ACT_1)
n2<-length(ACT_2)
data_2<-data.frame(data, fit$residuals)
resi_1<-data_2[ACT<26,]$fit.residuals
resi_2<-data_2[ACT>=26,]$fit.residuals
m_1<-median(resi_1)
m_2<-median(resi_2)
d1<-abs(resi_1-m_1)
d2<-abs(resi_2-m_2)
d1_m<-mean(d1)
d2_m<-mean(d2)
s_power<-(sum((d1-d1_m)^2)+sum((d2-d2_m)^2))/(n1+n2-2)
t_bf<-(d1_m-d2_m)/(sqrt(s_power)*sqrt(1/n1+1/n2))
```

The result is $|t^*_{BF}| = 0.8967448 < 2.61814$; Therefore, conclude H0. The error variance is constant (error variance is independent with X). And yes, the result here is consistent with my preliminary finding in part c that errors are independent and the assumptions of linear regression are satisfied.

Part f

```
data_3<-read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/tex
names(data_3)<-c("GPA","ACT","Intelligence","Rank")
attach(data_3)
plot(Intelligence,resi)
plot(Rank,resi)
```



I can see the residual has systematic positive trend with the intelligence test scores, but no relationship with the high school rank. So we should only add the intelligence test score variable in the regression model, which has the potential to improve our original regression line based on ACT score only.

3.15

Part a

```
data<-read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassno
names(data)<-c("con","time")
attach(data)
fit<-lm(con~time)
summary(fit)
```

```
Call:
lm(formula = con ~ time)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5333 -0.4043 -0.1373  0.4157  0.8487 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.5753     0.2487  10.354 1.20e-07 ***
time         -0.3240     0.0433  -7.483 4.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4743 on 13 degrees of freedom
Multiple R-squared:  0.8116,    Adjusted R-squared:  0.7971 
F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06
```

bo = 2.57533, b1 = - 0.32400

The regression function is Y_hat = 2.57533-0.32400X

Part b

Ho: E [Y] = $\beta_0 + \beta_1 X$
H1: E [Y] $\neq \beta_0 + \beta_1 X$

$\alpha = 0.025$
Decision Rule: If $F^* <= F(1-\alpha, c-2, n-c) = F(.975, 3, 10) = 4.83$, conclude Ho.Otherwise, reject Ho and conclude that the regression function does not adequate fit the data (a significant lack of fit exists in the linear model)

Code:

```
# part b
reduced<-lm(con~time)
full<-lm(con~0+as.factor(time))
anova(reduced, full)
```

```
Analysis of Variance Table

Model 1: con ~ time
Model 2: con ~ 0 + as.factor(time)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     13 2.9247
2     10 0.1574  3    2.7673 58.603 1.194e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$F^* = 58.603 > 4.83$, reject Ho and we conclude there exists a lack of fit.

Part c

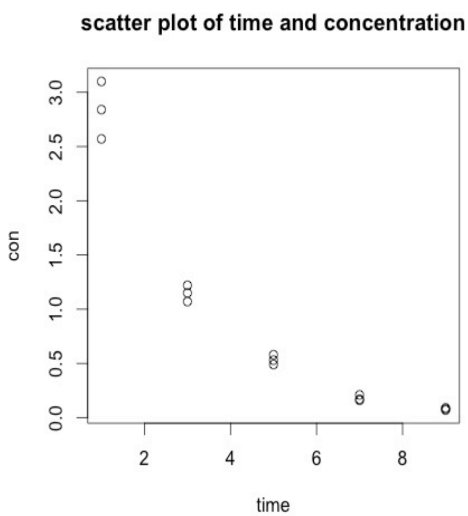
The lack of fit test indicates that the regression function is not linear. This means that regression function must be non-linear, for example, quadratic.

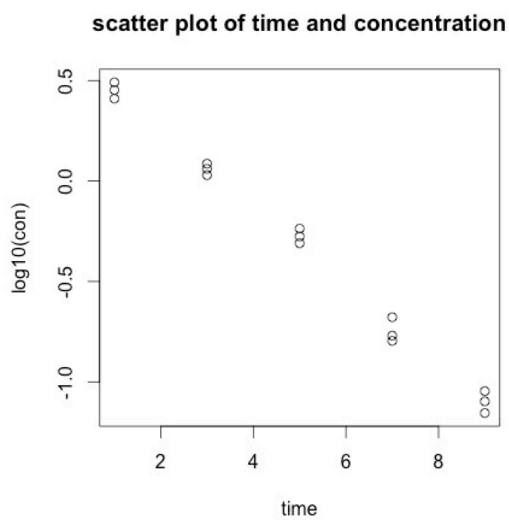
3.16

Part a

```
# part a
plot(time,con,main="scatter plot of time and concentration")
plot(time,log10(con),main = "scatter plot of time and concentration")
```

The original scatter plot of x and y is displayed on the left; the scatter plot after the transformation of Y is displayed on the right.





According to the prototype in figure 3.15, we use $Y' = \log_{10} Y$ because the scatter plot matches the prototype (b), where the variance of the error is larger (or we say the variance of Y is large) when x is small.

Part c

Code:

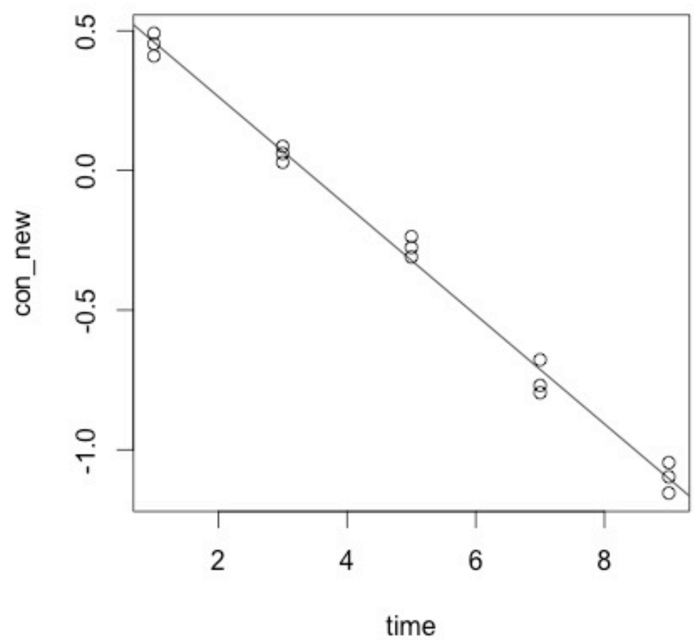
```
# part c
con_new<-log10(con)
fit_new<-lm(con_new~time)
summary(fit_new)
```

```
Call:
lm(formula = con_new ~ time)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.082958 -0.044421  0.006813  0.033512  0.085550 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.654880   0.026181   25.01 2.22e-12 ***
time        -0.195400   0.004557  -42.88 2.19e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04992 on 13 degrees of freedom
Multiple R-squared:  0.993,	Adjusted R-squared:  0.9924 
F-statistic:  1838 on 1 and 13 DF,  p-value: 2.188e-15
```

Regression function is Y’_hat = 0.654880 - 0.195400X
with b0=0.654880,b1=0.195400

Part d

```
# part d
plot(time,con_new)
abline(lm(con_new~time))
```

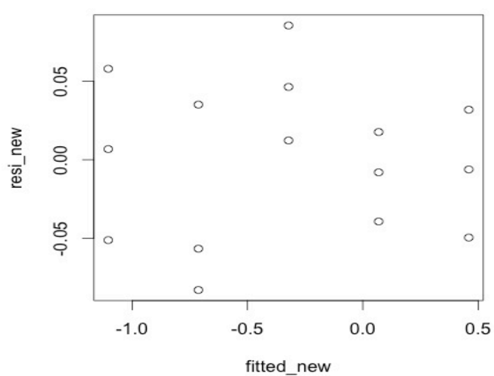


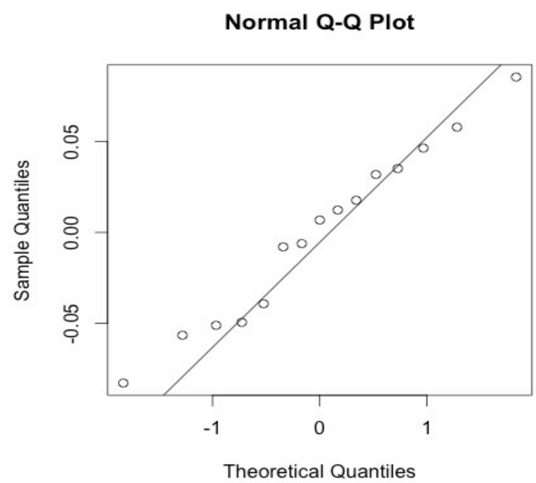
It looks like the estimated regression line Y'_hat = 0.65488−0.19540X fits the transformed data for Y' = log10 Y well.

Part e

```
# part e
resi_new<-fit_new$residuals
fitted_new<-fit_new$fitted.values
plot(fitted_new,resi_new)
qqplot<-qqnorm(resi_new)
qqline(resi_new)
```

```
> fit_new$residuals
           1            2            3            4            5            6            7            8 
-0.051178946  0.057965523  0.006813001 -0.082957620 -0.056628681  0.035141692  0.012317861  0.085549775 
           9           10           11           12           13           14           15 
 0.046397651  0.017680995 -0.007980995 -0.039295058 -0.006161112 -0.049546328  0.031882242 
> 
```





The residual plot shows that the residuals has fairly constant variability along the fitted the value, which meets the assumption of linear regression model that the error is independent variable; and the normal probability plot shows the residuals are approximately normally distributed since it is fairly linear.