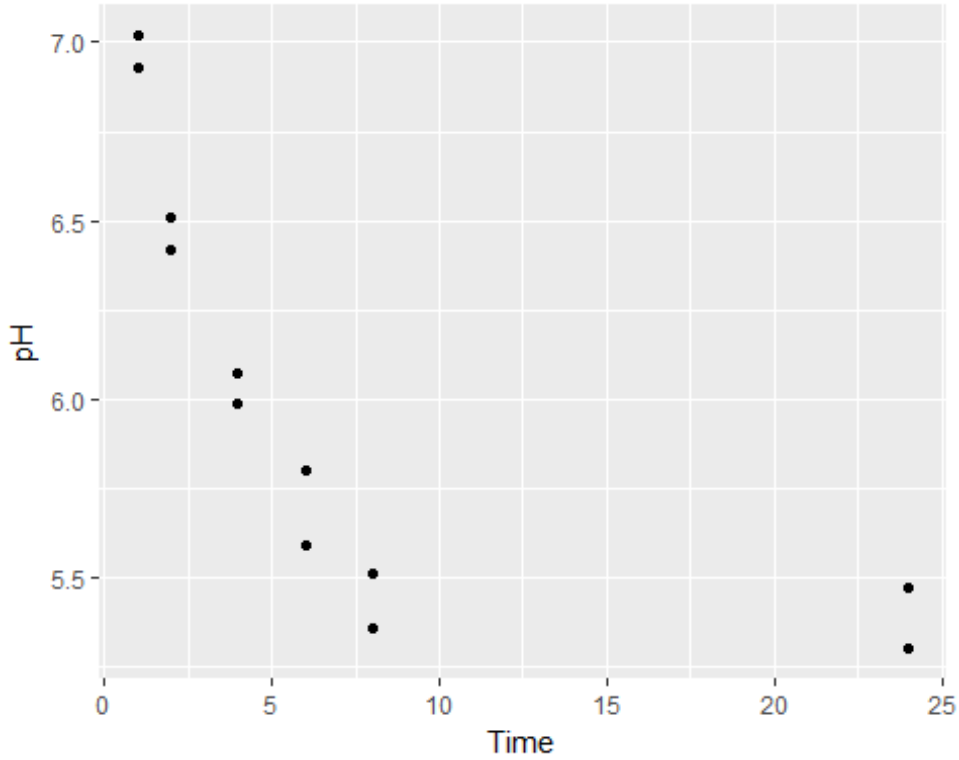
Assignment-8.R

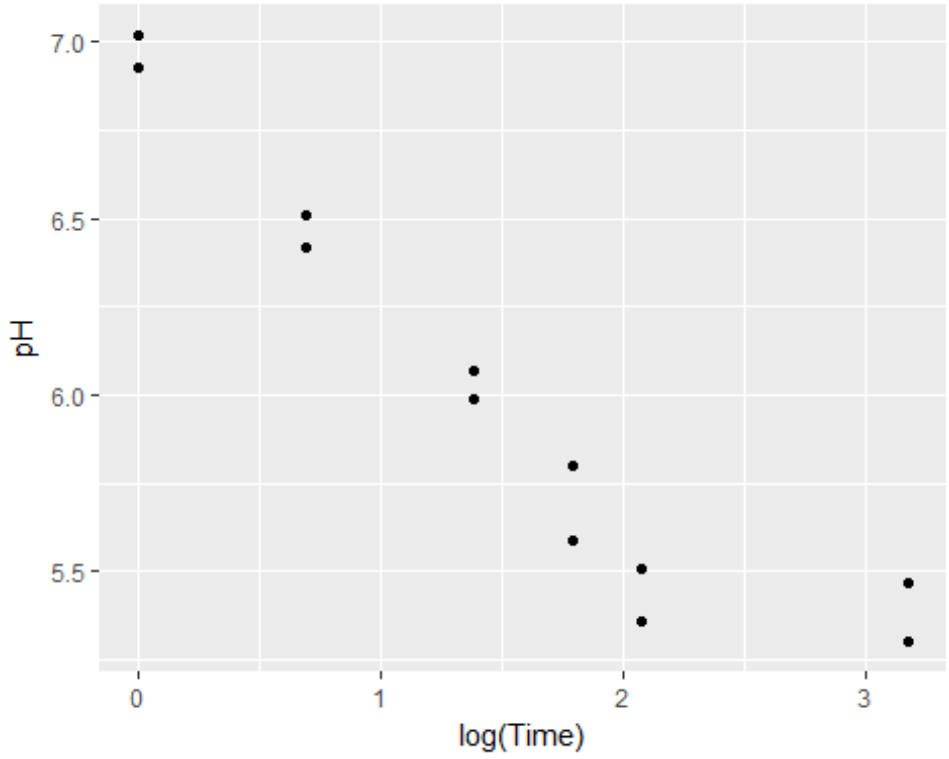
Purbasha Chatterjee

Thu Nov 30 15:12:27 2017

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.2

library(Sleuth3)

## Warning: package 'Sleuth3' was built under R version 3.4.2

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.4.2

## Q1-Ex:8.16
ggplot(data = ex0816, aes(Time, pH)) + geom_point()
```




```
ggplot(data = ex0816, aes(log(Time), pH)) + geom_point()
```



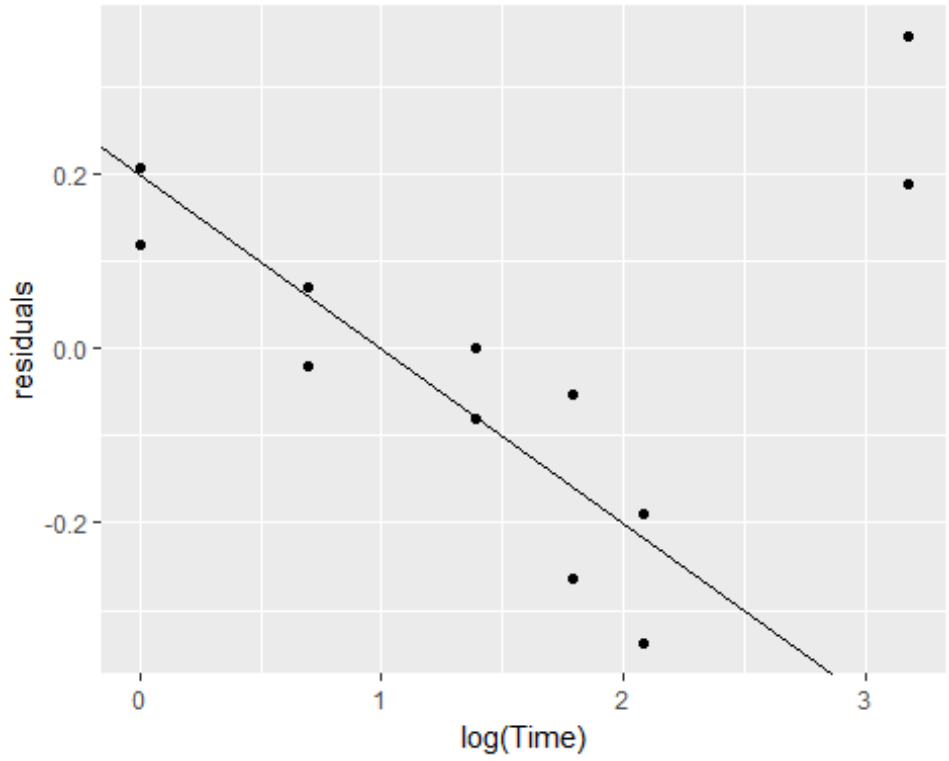

```
slr <- lm(pH ~ log(Time), data = ex0816)
summary(slr)
```

```
##
## Call:
## lm(formula = pH ~ log(Time), data = ex0816)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33897 -0.10710 -0.01023  0.13609  0.35879
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.8115     0.1113  61.205 3.30e-14 ***
## log(Time)    -0.5350     0.0609  -8.785 5.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2135 on 10 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8738
## F-statistic: 77.18 on 1 and 10 DF,  p-value: 5.14e-06
```

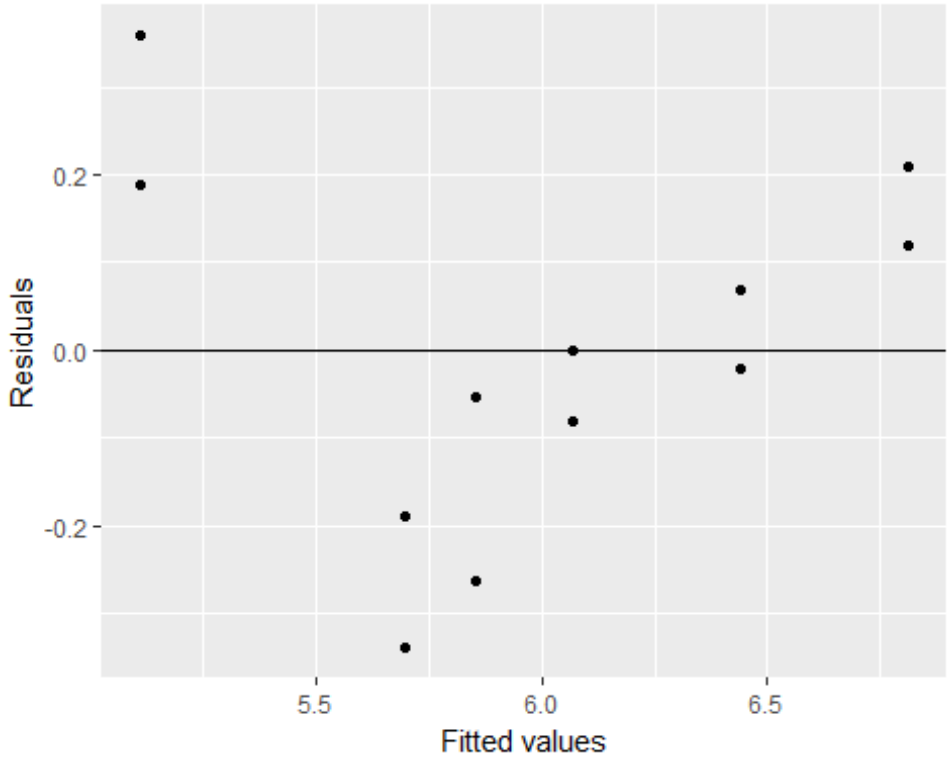
```
confint(slr)
```

```
##                  2.5 %     97.5 %
## (Intercept)  6.5635117  7.0594454
## log(Time)   -0.6706944 -0.3993097
```

```r
pHdata <- ex0816
fits <- predict(slr, se.fit = TRUE)
n <- dim(ex0816)[1]
M <- sqrt(2*qf(0.95, 2, n-2))
pHdata$lower <- fits$fit - M * fits$se.fit
pHdata$upper <- fits$fit + M * fits$se.fit
(bhat <- coefficients(slr))

## (Intercept)   log(Time)
##   6.8114785  -0.5350021

bhat <- as.numeric(bhat)
(estimate <- (6.0 - bhat[1])/bhat[2])

## [1] 1.516777

exp(estimate)

## [1] 4.55751

pHdata$residuals <- residuals(slr)

ggplot(data = pHdata, aes(log(Time), residuals)) + geom_point()+ geom_abline(
intercept = 0.2, slope = -0.2)
```




```r
qplot(slr$fitted.values, slr$residuals)+ geom_hline(aes(yintercept=0))+xlab("
Fitted values") + ylab("Residuals")
```




```
smm <- aov(pH ~ as.factor(log(Time)), data = pHdata)
summary(smm)

##                       Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(log(Time))  5  3.916  0.7833   79.59 2.11e-05 ***
## Residuals             6  0.059  0.0098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(slr, smm)

## Analysis of Variance Table
##
## Model 1: pH ~ log(Time)
## Model 2: pH ~ as.factor(log(Time))
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     10 0.45602
## 2      6 0.05905  4   0.39697 10.084 0.007841 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) In fitted regression, the intercept estimate is observed to be 6.81 giving a mean of 6.81-o.535*log(Time). By plotting residual against log(Time); it can be observed that with the increase in log(Time), the residuals decrease except for the last 2 data points. When the fitted values are assessed against residual with the help of residual plot, it could be observed that two data points lie extremely away from y-intercept 0

and even from the range of 0.2 t0 -0.2 y-intercept (although 1 more point is slight away from -0.2). Additionally, in the fitted line of residual against log(Time), all the points do not fit well to line with 2 extreme points at the end. This raises the doubt that 2 data points are the outliers in these data. It shows a curvature.

b) From the F-test it could be observed that p-value is less than 0.05, this means we can reject the null hypothesis. This denotes that separate mean model is no better than regression model and thus a lack of fit exist on the straight line.

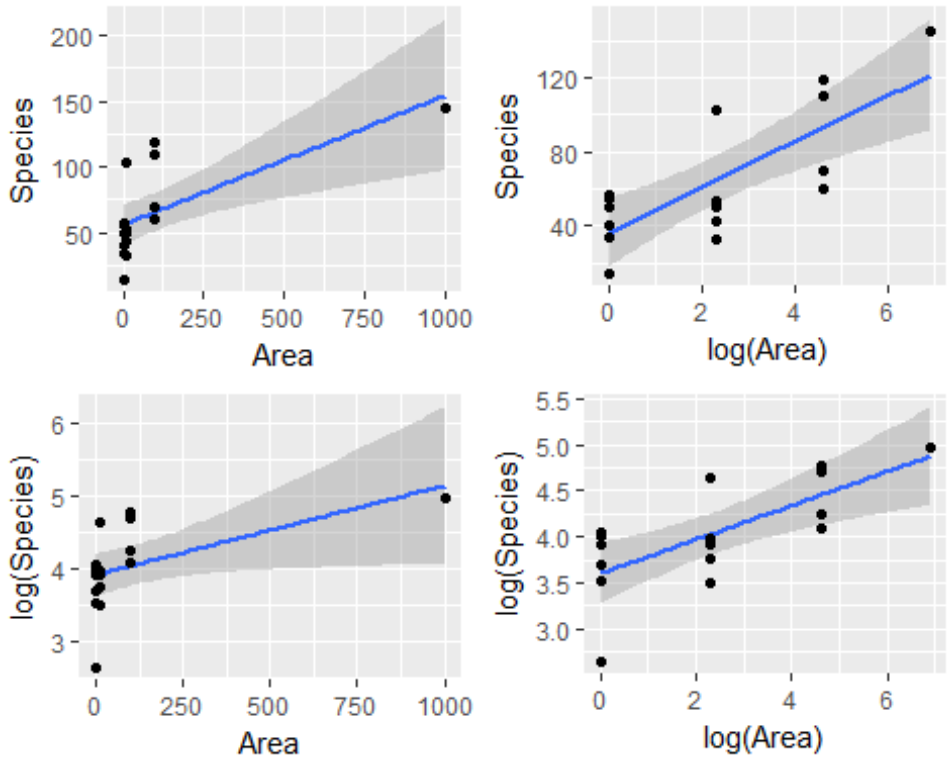
c) From the data and findings, it could be observed that straight line fit does not work over it properly. The points between 1 to 8 should not be dropped but dropping the outliers might help to get the proper mean value, because after adding the last 2 data points, the pH mean value increased from 3.8 to 4.5 hours.

Additional observations were made:

After applying linear regression, slope estimate was observed to be -0.535. The population's standard point was observed to be 0.2135 with variance as 0.8853. Additionally, the F-statistics value was observed to be 77.18 and p-value is less than 0.05 allowing to reject null hypothesis. The 95% confidence interval for the slope was observed to be 0.3993097 to 0.6706944. From the Separate mean model, it could be seen that full model supports over reduced model.

Hence, it could be deduced that the pH decreases with the increase in log of time(hours). Although with some extreme data-points, it could be said that pH of 6 could be achieved at somewhat around 4^{th} hour.

```
## Q2-Ex:8.22

p1<- ggplot(data = ex0822, aes(Area, Species)) + geom_smooth(method = "lm") +
geom_point()
p2<- ggplot(data = ex0822, aes(log(Area), Species)) + geom_smooth(method = "l
m") + geom_point()
p3<- ggplot(data = ex0822, aes(Area, log(Species))) + geom_smooth(method = "l
m") + geom_point()
p4<- ggplot(data = ex0822, aes(log(Area), log(Species))) + geom_smooth(method
= "lm") + geom_point()
grid.arrange(p1,p2,p3,p4, ncol = 2)
```




```
slr1 <- lm(Species ~ log(Area), data = ex0822)
summary(slr1)

##
## Call:
## lm(formula = Species ~ log(Area), data = ex0822)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33.25 -21.88   0.75  19.25  38.25
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.250      8.702   4.166 0.000952 ***
## log(Area)     12.377      2.760   4.485 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.78 on 14 degrees of freedom
## Multiple R-squared:  0.5896, Adjusted R-squared:  0.5603
## F-statistic: 20.11 on 1 and 14 DF,  p-value: 0.000514

confint(slr1)

##                 2.5 %   97.5 %
## (Intercept) 17.586380 54.91362
## log(Area)    6.457967 18.29682
```

```r
smm1 <- aov(Species ~ as.factor(log(Area)), data = ex0822)
anova(slr1, smm1)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ log(Area)
## Model 2: Species ~ as.factor(log(Area))
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     14 7915.5
## 2     12 6801.3  2    1114.2 0.9829 0.4024
```

```r
slr2 <- lm(log(Species) ~ log(Area), data = ex0822)
summary(slr2)
```

```
##
## Call:
## lm(formula = log(Species) ~ log(Area), data = ex0822)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9597 -0.2170  0.0180  0.3172  0.6103
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.59876    0.15407  23.359  1.3e-12 ***
## log(Area)    0.18486    0.04886   3.783  0.00202 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.421 on 14 degrees of freedom
## Multiple R-squared:  0.5055, Adjusted R-squared:  0.4702
## F-statistic: 14.31 on 1 and 14 DF,  p-value: 0.002017
```

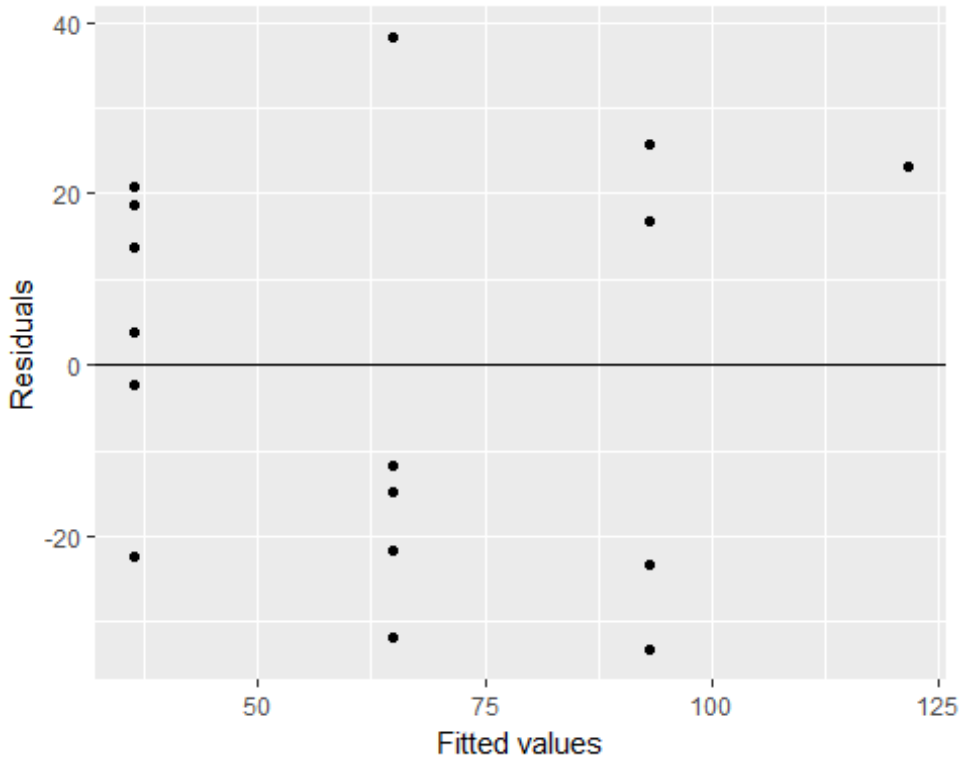
```r
confint(slr2)
```

```
##                  2.5 %    97.5 %
## (Intercept) 3.26831930 3.9291947
## log(Area)   0.08005785 0.2896636
```

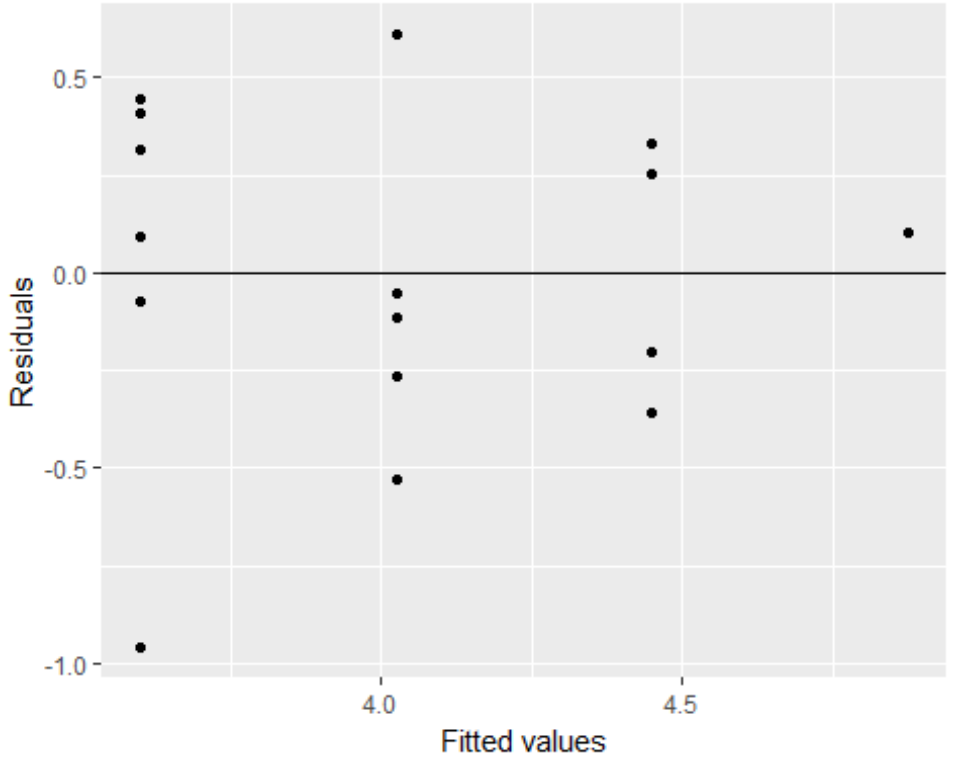
```r
smm2 <- aov(log(Species) ~ as.factor(log(Area)), data = ex0822)
anova(slr2, smm2)
```

```
## Analysis of Variance Table
##
## Model 1: log(Species) ~ log(Area)
## Model 2: log(Species) ~ as.factor(log(Area))
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     14 2.4812
## 2     12 2.4384  2  0.042826 0.1054 0.9008
```

```r
qplot(slr1$fitted.values, slr1$residuals)+ geom_hline(aes(yintercept=0))+xlab
("Fitted values") + ylab("Residuals")
```




```
qplot(slr2$fitted.values, slr2$residuals)+ geom_hline(aes(yintercept=0))+xlab
("Fitted values") + ylab("Residuals")
```



- From the above graph, it could be seen that regression model of Species against log(Area) and regression model of log(Species) against log(Area) captures most of the required confidence band. The other two graphs capture the widened confidence band with no data-points as most of the points are concentrated at the initial stage of the regression line, depicting a weak linear relation.
- For linear regression model of Species against log(Area), the slope estimate was found to be 12.377 with p-value as 0.000514. Thus, we can reject the null hypothesis and state that slope is not equal to 0. There exists a positive relation between the slope with intercept, giving a linear equation of 12.377X+36.25. The F-value is 20.11 on 15 df with population standard deviation as 23.78 and proportion of variance is 0.5896. The 95% confidence interval range is 6.458 to 18.297.
- Applying ANOVA over separate mean model and regression model gives a p-value of 0.402, failing to reject null hypothesis. Thus, lack of fit could not be stated.
- Except one data-point, the residual against fitted values remain in range -20 to 20 in the residual plot.
- For linear regression model of log(Species) against log(Area), the slope estimate was found to be 0.185 with p-value as 0.00207. Thus, we can reject the null hypothesis and state that slope is not equal to 0. There exists a positive relation between the slope with intercept, giving a linear equation of 0.185 X+3.599. The F-value is 14.31 on 15 df with population standard deviation as 0.421 and proportion of variance is 0.5055. The 95% confidence interval range is 0.0801 to 0.2897.
- Applying ANOVA over separate mean model and regression model gives a p-value of 0.9008, again failing to reject null hypothesis.
- Except one data-point, the residual against fitted values remain in range -1.0 to 0.5 in the residual plot.
- Thus, it was observed that, there was no significant difference in data fit for both the models and hence, it is not necessary to consider the log transformation of response. Additionally, the original response displayed more equality in variation in the scatter plot as compared to the transformed response in the plot.
- There exists a weak positive relative between the species and area, that is with every increase in log(Area), there exist a possibility of increase in species. But there exist, an outlier at point 1000 area which affects the equation as most of the other area points lies between 1-100 - affecting the regression line.