BTRY 6020 Lab II

##Which Glucose measure to Use?

The following exercise (which uses a real data set) demonstrates all the points we have discussed in class about simple regression. The following points are covered:

1) The usefulness of plotting both the data and the residuals to determine the fit of the regression model.
2) The need to pay special attention to outliers and influential points.
3) The use of transformations of the data to improve the fit and meet model assumptions.
4) Use of tests and confidence intervals when the data have been transformed.

Problem: The concentration of glucose in the blood can be measured using a colorimeter and a substance that changes color in the presence of glucose. This test is highly sensitive, but is very vulnerable to contamination of the sample. Another test uses an enzyme, which is specific to glucose. This test is less vulnerable to contamination, has been used for years, and is highly accurate. The colorimetric test is less than one-tenth the cost of the other, and administrators wonder if they can switch over to the cheaper test. To examine the relationship between the results of these two tests, blood samples were taken from 46 patients at Stanford Hospital. The glucose concentration in each sample was measured using both methods.

The main objective of the study is to determine how well the colorimetric values predict the enzymatic values. A prediction (regression) equation will be fitted, and the fit of the model examined.

A second objective is to determine a confidence interval for the mean enzymatic value of patients with colorimetric value 240, and a prediction interval for new patients with this colorimetric value.

The analysis begins by loading and plotting the data, and fitting the regression equation.

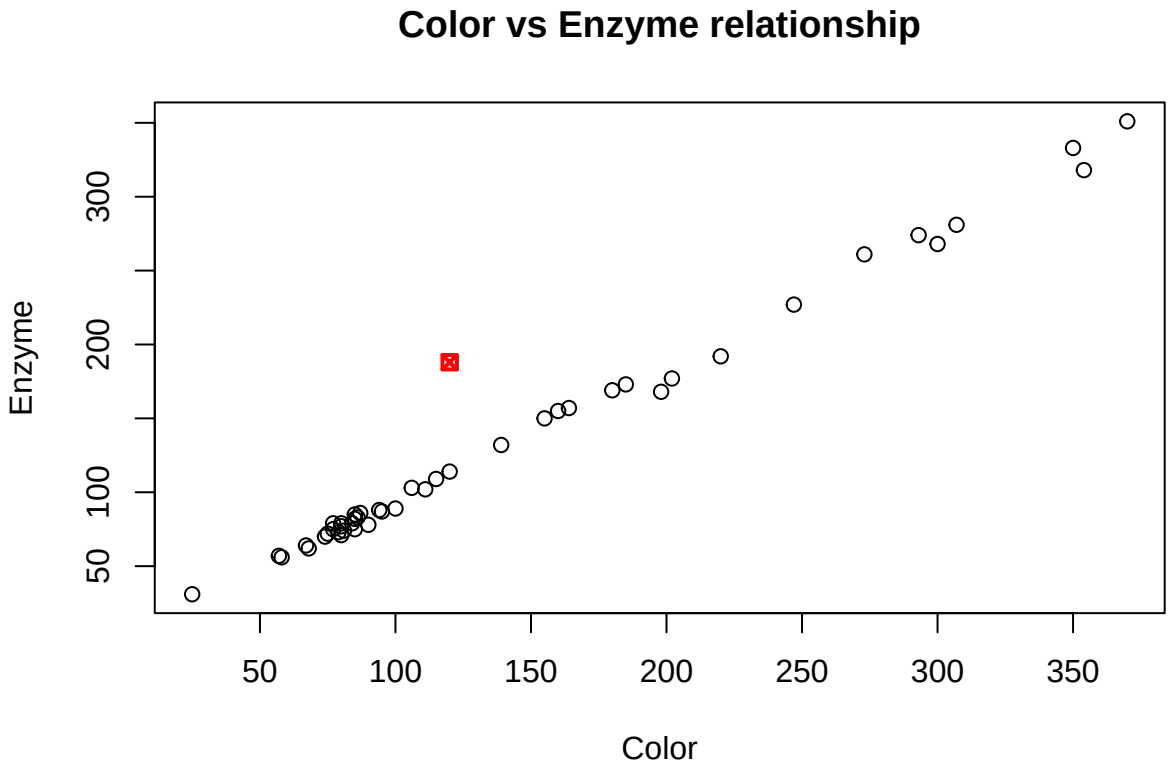
The data listed below are the two measures of glucose from blood samples taken from 46 patients at Stanford Medical Center in 1976.

```
library(readxl)
ColEnz = read_excel("ColEnz.xlsx")
head(ColEnz)
```

```
## # A tibble: 6 x 3
##   ObsNumber Color Enzyme
##       <dbl> <dbl>  <dbl>
## 1         1   155    150
## 2         2    80     79
## 3         3   139    132
## 4         4   293    274
## 5         5    67     64
## 6         6   354    318
```

The data are plotted below

```
plot(ColEnz$Color, ColEnz$Enzyme, xlab = "Color", ylab = "Enzyme", main = "Color vs Enzyme relationship
points(ColEnz$Color[16], ColEnz$Enzyme[16], lwd = 2, col = "red", pch = 7)
```



Based on the following linear model summary, answer the following questions.

```
ColEnz.lm = lm(Enzyme~Color, data = ColEnz)
summary(ColEnz.lm)
```

```
##
## Call:
## lm(formula = Enzyme ~ Color, data = ColEnz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.006  -4.260  -0.646   1.831  73.900
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.01364    3.47318   1.444    0.156
## Color        0.90905    0.02068  43.962   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.55 on 44 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9772
## F-statistic:  1933 on 1 and 44 DF,  p-value: < 2.2e-16
```

a) What is the value of R^2? Does this indicate a good fit to the line?

b) Since both laboratory methods measure the same thing, the regression equation should have the form : $y = x + \epsilon$

That is, $\beta_0 = 0$ and $\beta_1 = 1$. Do the data support this? (Do not worry about multiple comparisons here.)

```r
# hypothesis testing for intercep: HO: BO=0; HA: BO not equal to zero
tstat0 = (5.01364 - 0) /(3.47168)
tstat0
```

```
## [1] 1.444154
```

```r
p_value0= 2*(1-pt(abs(tstat0), df=44))
p_value0
```

```
## [1] 0.1557792
```

```r
p_value0<0.05
```

```
## [1] FALSE
```

We fail to reject H0, about that BO=0.

```r
# hypothesis testing for slope: HO: B1=1; HA: B1 not equal to one
tstat1 = (0.90905 - 1) /(0.02068)
tstat1
```

```
## [1] -4.397969
```

```r
p_value1 = 2*(1-pt(abs(tstat1), df=44))
p_value1
```

```
## [1] 6.833957e-05
```

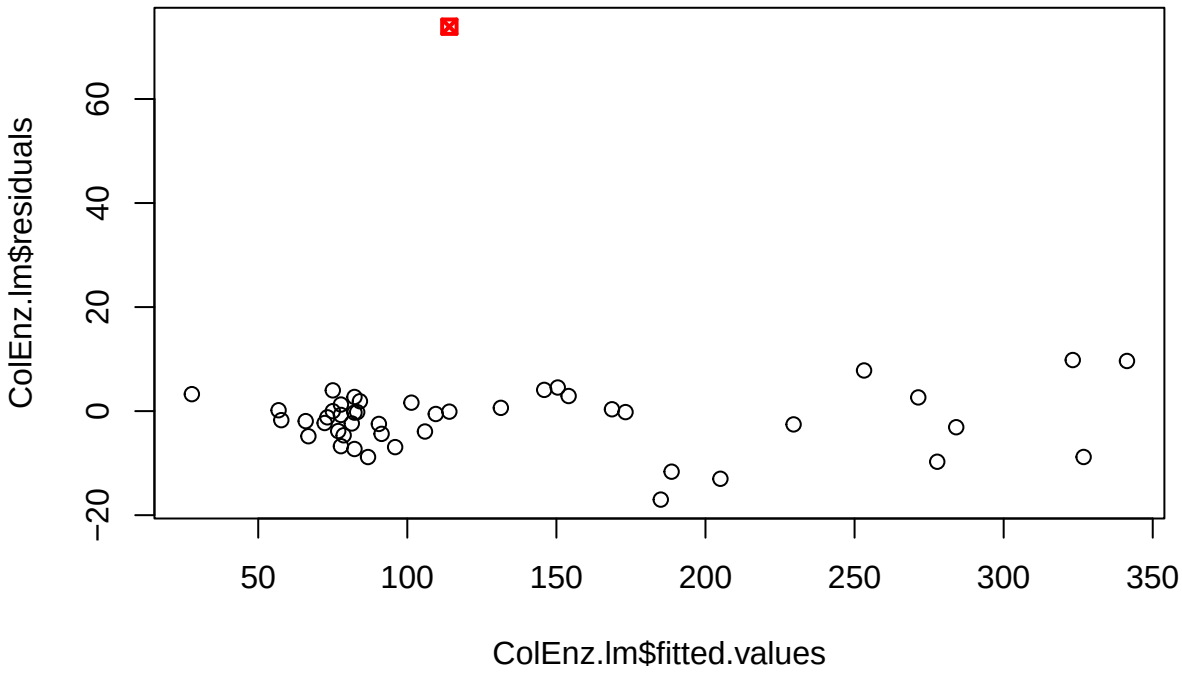
```r
p_value1<0.5
```

```
## [1] TRUE
```

At 0.05 significance level, we reject the null hypothesis that the null=1.

c) The plot of residuals from this model against the predicted values is below. Does this plot have any statistically interesting features?

```r
plot(ColEnz.lm$fitted.values, ColEnz.lm$residuals)
points(ColEnz.lm$fitted.values[16], ColEnz.lm$residuals[16], lwd = 2, col = "red", pch = 7)
```



d) The isolated point on this plot is patient 16. Patient 16 is also labeled on the scatterplot of the raw data. What should be done about this value?

The red point is a outlier, this patien was bad labelled.

e) Actually, patient 16 was entered incorrectly into the computer. The colorimetric value of this patient should be 210, but was entered as 120. Below is the edited data set and the new linear regression fit. What effect did the error have on the regression equation? Why

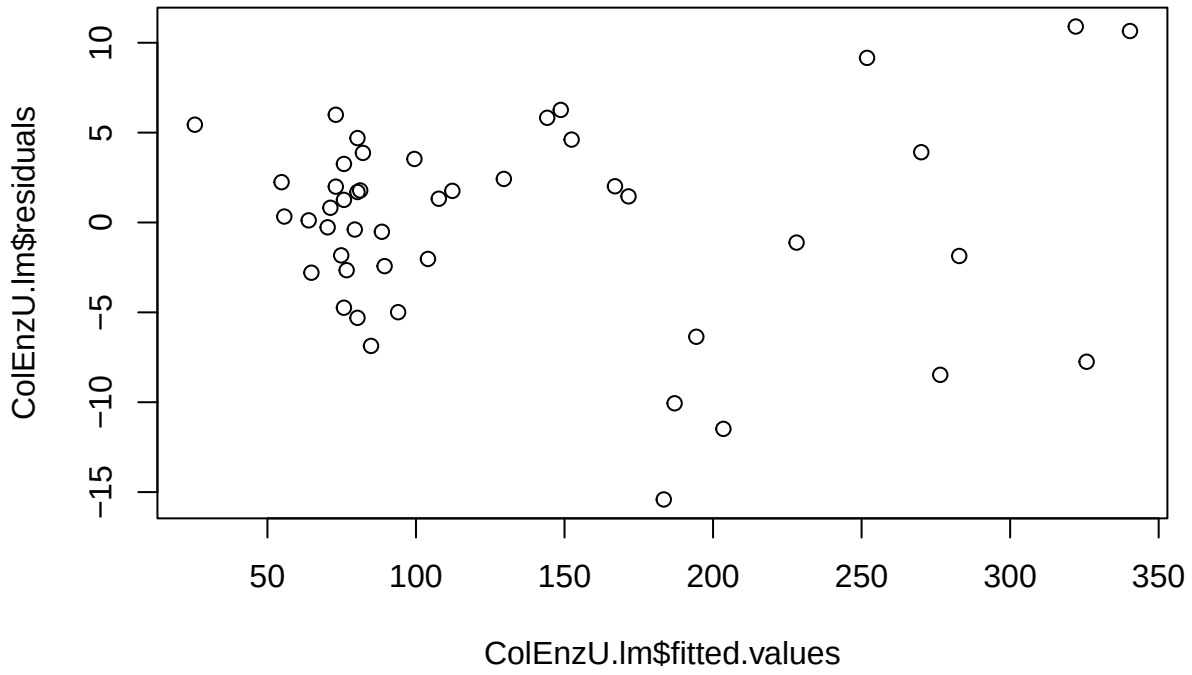
```
#save new data set
ColEnzU = ColEnz
#edit the colorimetric value of the 16th patient
ColEnzU[16,2] = 210

#fit the new regression
ColEnzU.lm = lm(Enzyme~Color, data = ColEnzU)
summary(ColEnzU.lm)
```

```
##
## Call:
## lm(formula = Enzyme ~ Color, data = ColEnzU)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.409  -2.599   1.038   3.465  10.902
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.748267   1.561592    1.76   0.0854 .
## Color       0.912429   0.009192   99.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.61 on 44 degrees of freedom
```

```
## Multiple R-squared:  0.9956, Adjusted R-squared:  0.9955
## F-statistic:  9852 on 1 and 44 DF,  p-value: < 2.2e-16
```

f) What are the interesting features of the plot of residuals versus predicted values? What effect will this have on the estimate of the variance of β_0 vs. β_1? What effect will this have on the estimate of a confidence interval for the mean enzymatic value when the colorimetric value is 70? When it is 280?

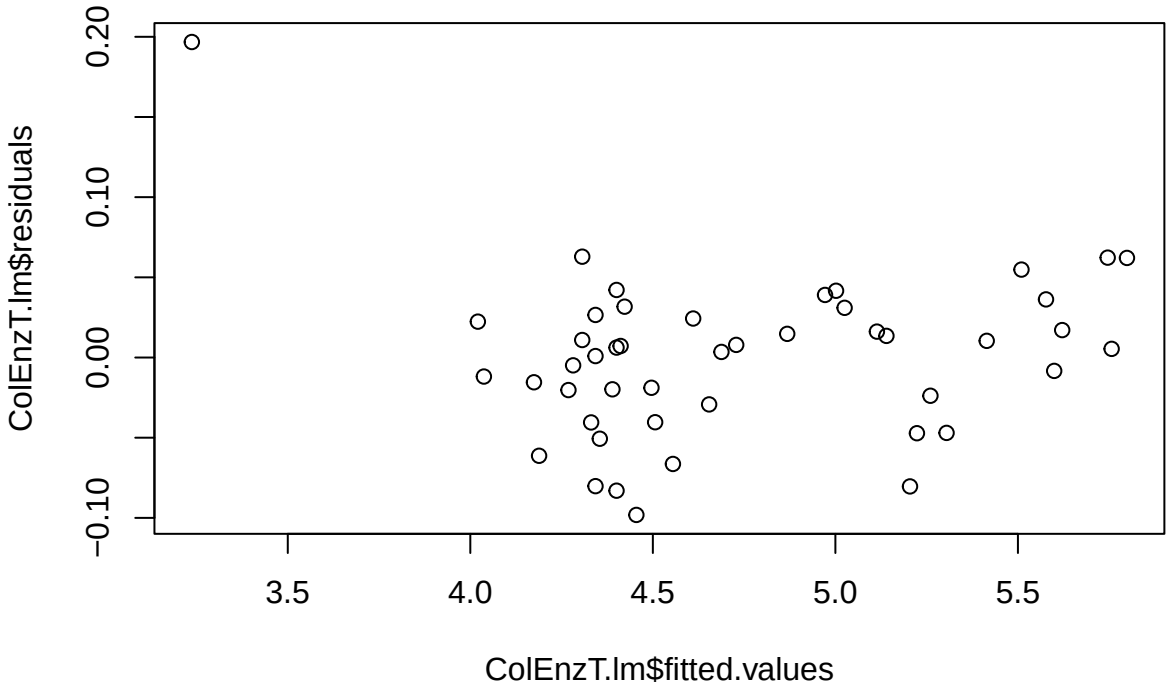
```
plot(ColEnzU.lm$fitted.values, ColEnzU.lm$residuals)
```



g) Suppose we consider the model $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i$. What do we expect the values of β_0 and β_1 to be?

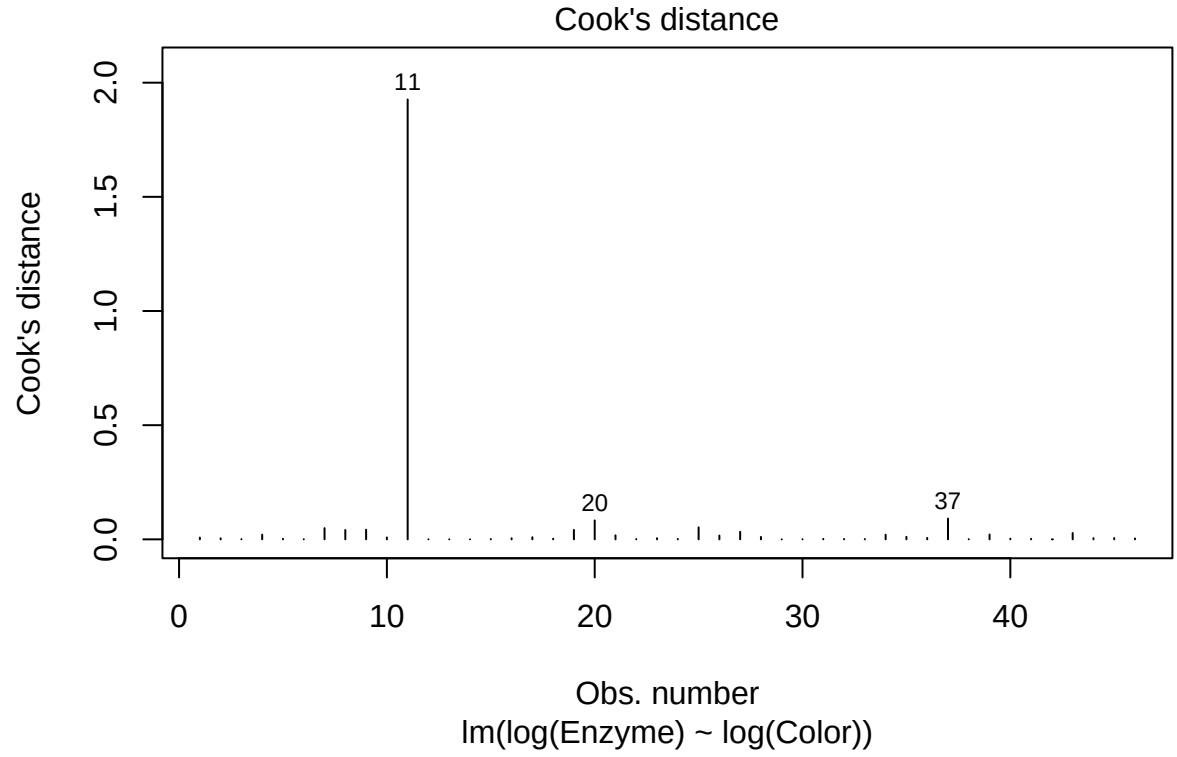
h) The plot of residuals versus predicted values for this model are below. What are the interesting features of this plot?

```
#log transformed linear regression
ColEnzT.lm = lm(log(Enzyme)~log(Color), data = ColEnzU)
#plotting  residuals vs predicted values
plot(ColEnzT.lm$fitted.values, ColEnzT.lm$residuals)
```

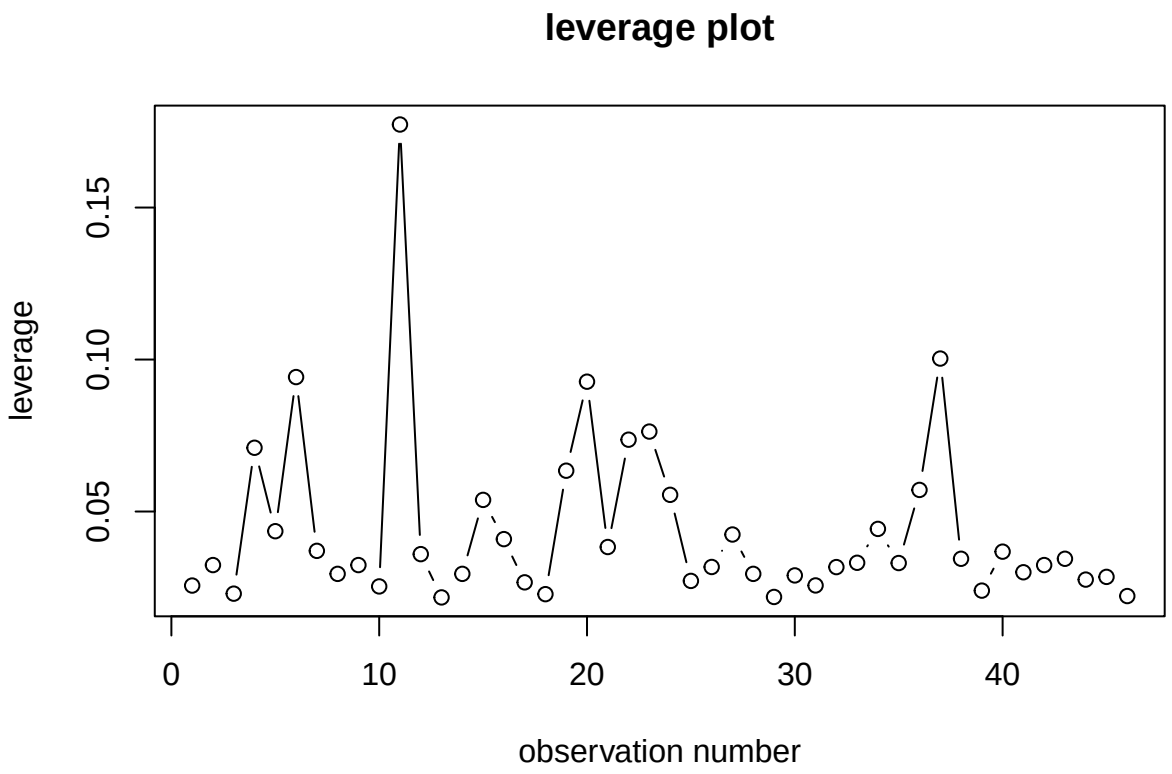


i) The following plots shows the influence (Cook's distance) and leverage of our observations. What are the interesting features of these plots?

```
#Cook's Distance
plot(ColEnzT.lm, which = 4)
```




```
#Leverage Plot
lev = hat(model.matrix(ColEnzT.lm))
plot(lev, type="b", xlab="observation number", ylab="leverage", main="leverage plot")
```



j) The isolated point on the residual plot is patient 11. Patient 11 has the highest leverage and Cook's distance. What is the effect of this data point on the regression equation? Checking against the patient chart shows that the correct data value was entered. Should we make any adjustment for this outlier?

k) The investigators decided to fit the regression equation without patient 11. The following is the linear model summary.

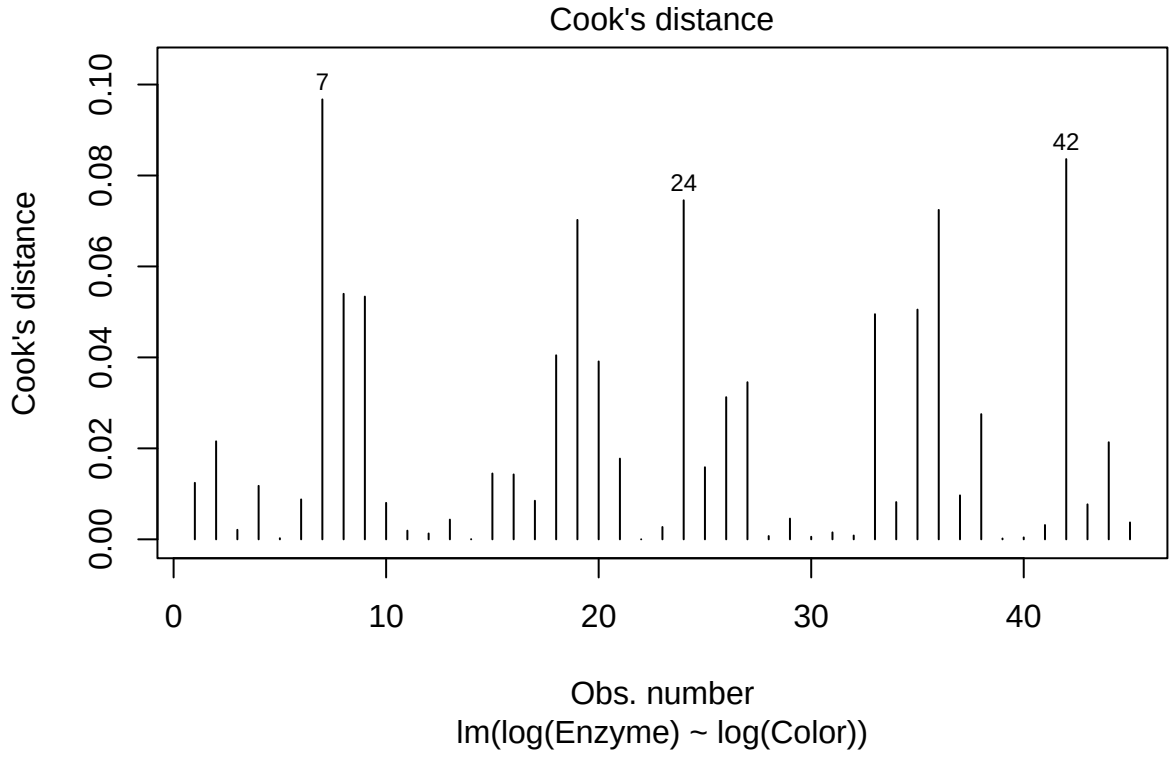
```
#remove 11th observations
ColEnzR = ColEnzU[-11,]

#fit linear model
ColEnzR.lm = lm(log(Enzyme)~log(Color), data = ColEnzR)
summary(ColEnzR.lm)
```

```
##
## Call:
## lm(formula = log(Enzyme) ~ log(Color), data = ColEnzR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08686 -0.02525  0.00984  0.02896  0.07872
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.05901    0.05306   1.112    0.272
## log(Color)   0.97420    0.01092  89.224   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

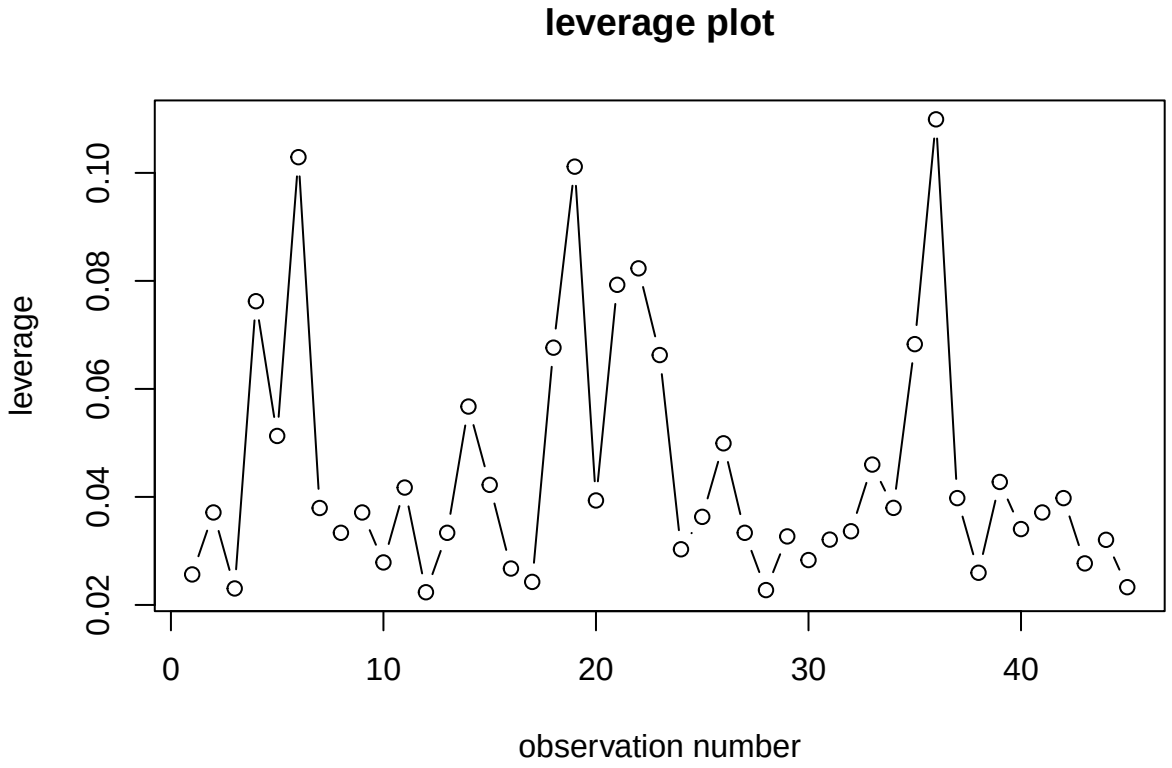
```
## Residual standard error: 0.03999 on 43 degrees of freedom
## Multiple R-squared:  0.9946, Adjusted R-squared:  0.9945
## F-statistic:  7961 on 1 and 43 DF,  p-value: < 2.2e-16
```

l) The residual plot for this model now looks fine. Below are the leverage and influence plots. Are there any further problems of high leverage or influence for this model?

```
#Cook's Distance
plot(ColEnzR.lm, which = 4)
```




```
#Leverage Plot
lev2 = hat(model.matrix(ColEnzR.lm))
plot(lev2, type="b", xlab="observation number", ylab="leverage", main="leverage plot")
```



m) What is the estimate of the regression equation? What is a 95% confidence interval for β_0? For β_1?

n) What is a 95% confidence interval for the mean enzymatic value when the colorimetric value is 240? ($S_{xx} = 13.41, \bar{x} = 4.83$) Please interpret.

o) What is a 95% prediction interval for the enzymatic value of a patient whose colorimetric value is 240? Please interpret.