36401 HW4

Sylvia

9/23/2019

Part A

```
BloodLevel = read.table("BloodLevel.dat",header = TRUE)
```

(a)

(1) **EDA**

```
plot(BloodLevel$Level~BloodLevel$Age, main = "Loess plot: Age vs Level", xlab = "Age", ylab = "Level")
lines(lowess(BloodLevel$Level~BloodLevel$Age, f = 8/10), lty = 2)
```

The trend seems to be like as Age increases for infants, the levels of hormone in the blood decreases. In other words, as x increases, y decreases. Currently from the graph, it is hard to tell if there is any outlier. Also the variability of Y seems to decrease as X increases as well.

(2) **Simple Linear Regression Line**

```
plot(BloodLevel$Level~BloodLevel$Age, main = "Loess plot: Age vs Level", xlab = "Age", ylab = "Level")
lines(lowess(BloodLevel$Level~BloodLevel$Age, f = 8/10), lty = 2)
abline(lm(BloodLevel$Level~BloodLevel$Age), lty = 1)
```

(3)

Assessing fit of model

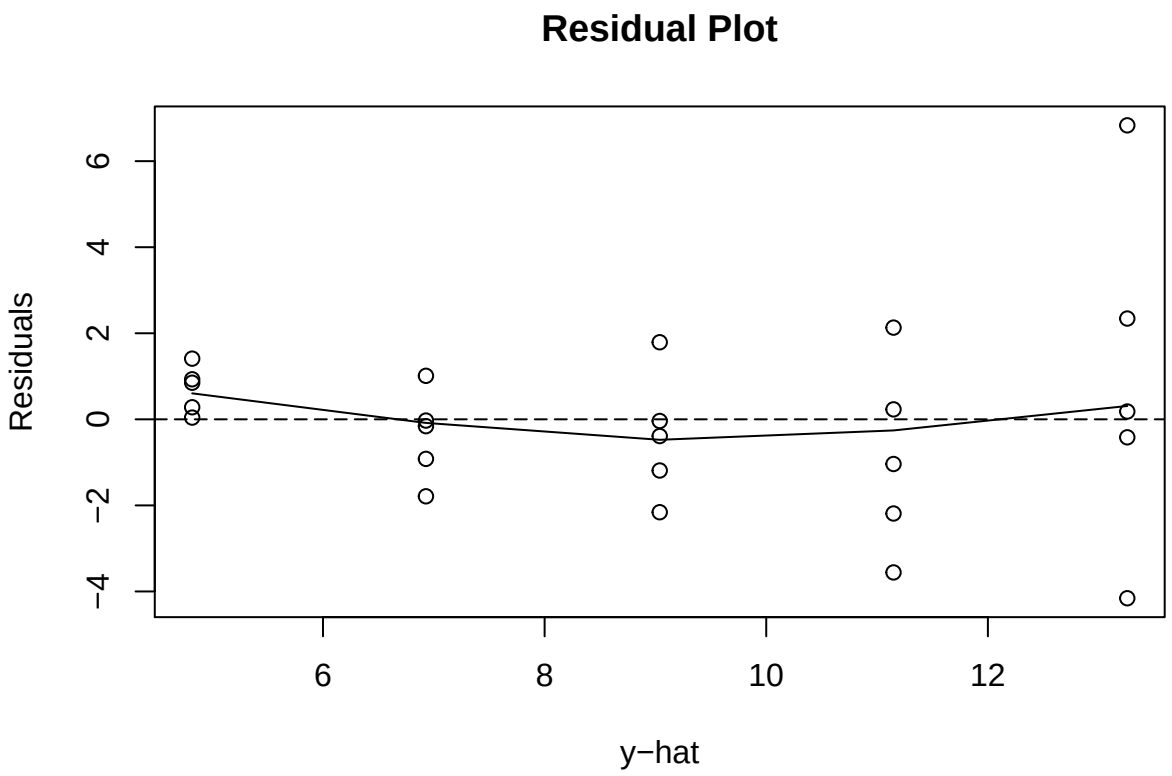
```
lm(BloodLevel$Level~BloodLevel$Age)
```

```
##
## Call:
## lm(formula = BloodLevel$Level ~ BloodLevel$Age)
##
## Coefficients:
##    (Intercept)  BloodLevel$Age
##          13.26           -2.11
```

The model seems to be a decent fit of the data, because here the R^2 value is 0.6672, which is realatively high in the range of [0,1], which means the proportion of variation in Y explained by the linear regression model is about 66%.

Residual Analysis

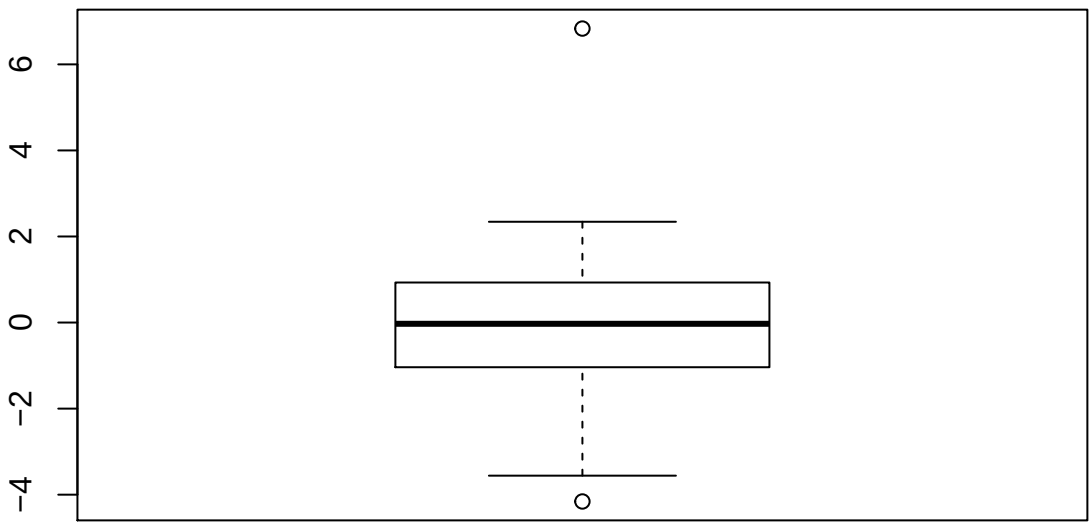
```
y.hat <- fitted(lm(BloodLevel$Level~BloodLevel$Age))
ep.hat <- resid(lm(BloodLevel$Level~BloodLevel$Age))
plot(y.hat, ep.hat, main = "Residual Plot", ylab = "Residuals", xlab = "y-hat")
abline(h=0, lty = 5)
lines(lowess(ep.hat~y.hat, f = 8/10, iter = 3), lty = 1)
```



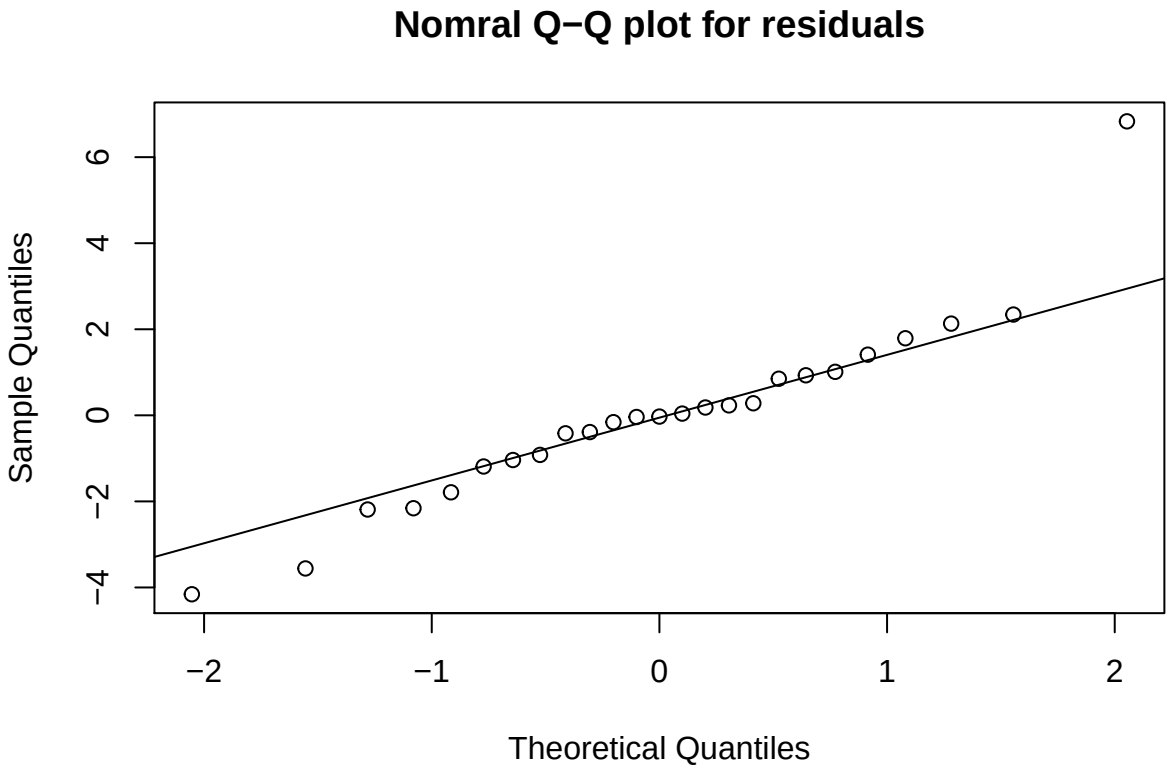

```
#legend(locator(n=1), legend=c("loess"), lty = 1)
```

Normality

```
boxplot(ep.hat, main = "Boxplot: Residuals")
```




```
qqnorm(ep.hat, main = "Nomral Q-Q plot for residuals")
qqline(ep.hat)
```



(4) **Test Significant Relationship**

```
summary(lm(BloodLevel$Level~BloodLevel$Age))
```

```
##
## Call:
## lm(formula = BloodLevel$Level ~ BloodLevel$Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1580 -1.0384 -0.0292  0.9304  6.8320
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     13.2580     0.7609  17.424 9.54e-15 ***
## BloodLevel$Age  -2.1096     0.3106  -6.791 6.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.197 on 23 degrees of freedom
## Multiple R-squared:  0.6672, Adjusted R-squared:  0.6528
## F-statistic: 46.12 on 1 and 23 DF,  p-value: 6.315e-07
```

Null Hypothesis: beta1 = 0, Alternative hypothesis: beta one != 0. The linear relaionship is significant because from the data we have, we can calculate the critical value of this T distribution. Since the estimated beta1 from above is -2.1096 and SE of estimated beta1 is 0.3106, t* = (-2.1096-0)/0.3106 = -6.792, and since the value is less than -t(1-alpha/2, df), which is 2.069, we reject the null hypothesis and conclude there is a statistically significant linear relationship between the two variables.

(b) Result: From the analysis, we see that there is a negative correlation between the variable Age and Level, and the linear relationship is significant because the p value less than 0.05. After fitting a linear

regression line, we have a model that has Intercept 13.258 and slope -2.1096, so we can predict the bloodlevel from age by the equation y = 13.258-2.1096t. From residual analysis, we can see that the residuals do not seem to be standard normally distributed and the normal Q-Q plot also indicates that there are outlier for the residual. Thus, we consider this model not really a good fit for the data. Also the QQ plot indicates normality assumption is violated here as well. Lastly, since we reject the null hypothesis that beta1 = 0 based on statistics given, we are confident that there is some negative correlation existent between variables age and level.

Part B

(c) Because the variabilities of residuals are not constant, and by looking at the concaved up feature of the lowess line, Level seems need a log transformation so that the trend can be corrected.

(d)

(1) **EDA**

```
log.level <- log(BloodLevel$Level)
plot(log.level~BloodLevel$Age, main = "Loess plot: Age vs log(Level)", xlab = "Age", ylab = "log(Level)
lines(lowess(log.level~BloodLevel$Age, f = 8/10), lty = 2)
```

(2) **Simple Linear Regression Line**

```
plot(log.level~BloodLevel$Age, main = "Loess plot: Age vs log(Level)", xlab = "Age", ylab = "log(Level)
lines(lowess(log.level~BloodLevel$Age, f = 8/10), lty = 2)
abline(lm(log.level~BloodLevel$Age), lty = 1)
```

(3)

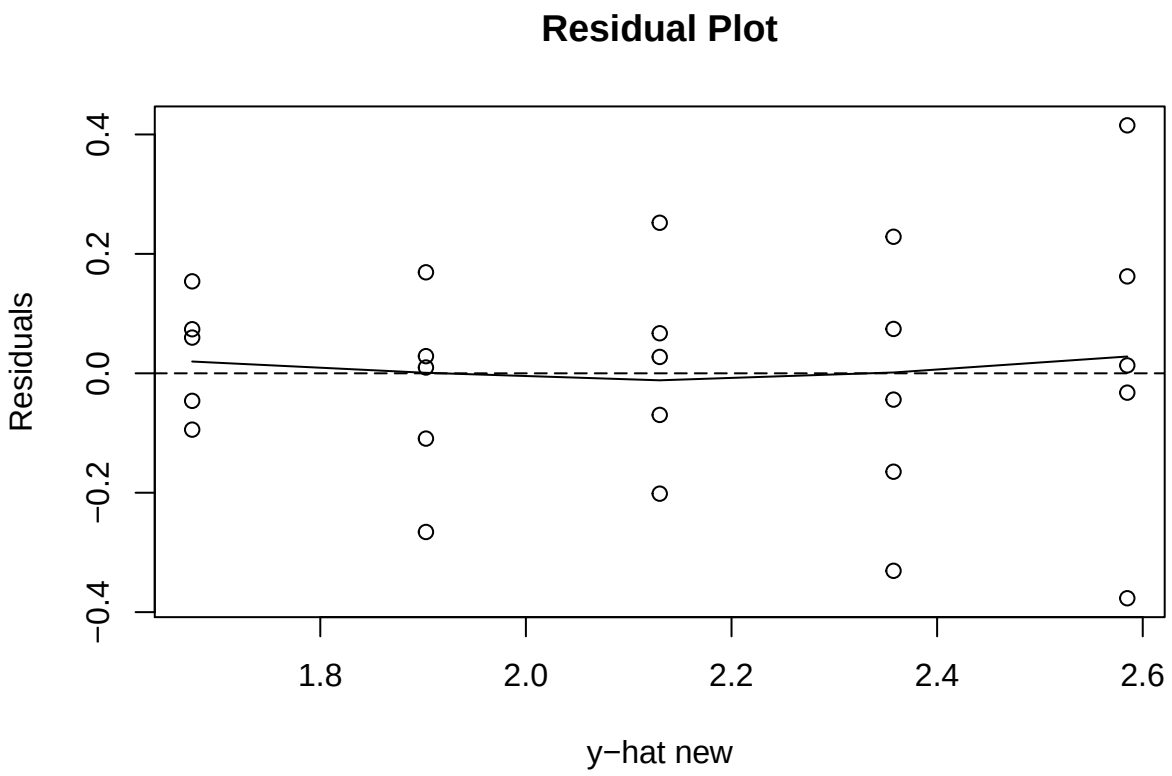
Assessing fit of model

```
lm(log.level~BloodLevel$Age)
```

```
##
## Call:
## lm(formula = log.level ~ BloodLevel$Age)
##
## Coefficients:
##    (Intercept)  BloodLevel$Age
##         2.5850         -0.2274
```

Residual Analysis

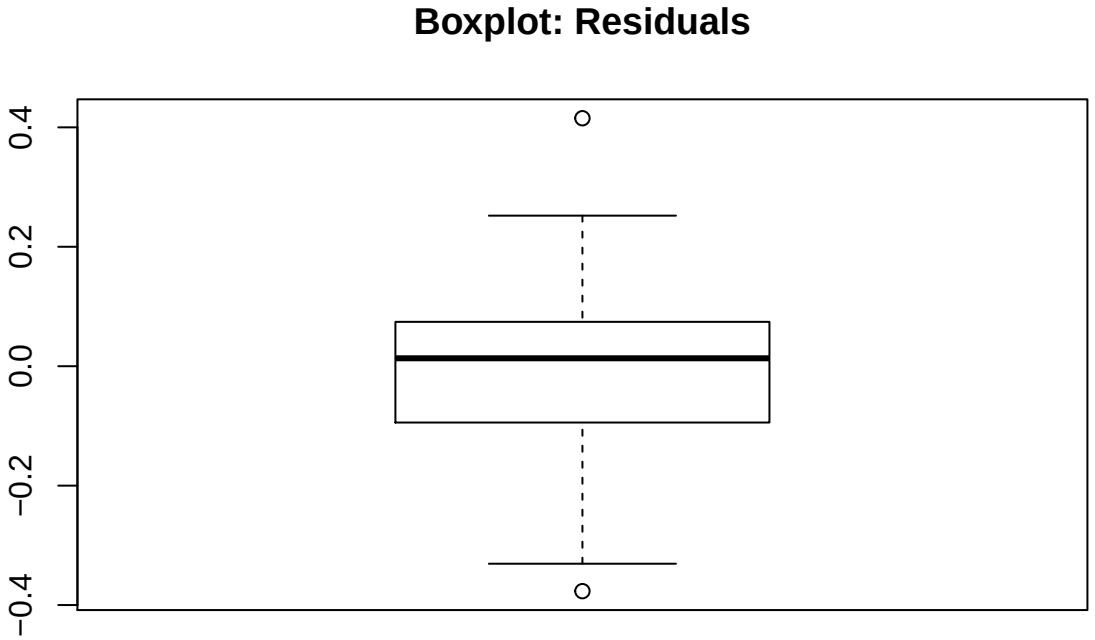
```
y.hat1 <- fitted(lm(log.level~BloodLevel$Age))
ep.hat1 <- resid(lm(log.level~BloodLevel$Age))
plot(y.hat1, ep.hat1, main = "Residual Plot", ylab = "Residuals", xlab = "y-hat new")
abline(h=0, lty = 5)
lines(lowess(ep.hat1~y.hat1, f = 8/10, iter = 3), lty = 1)
```



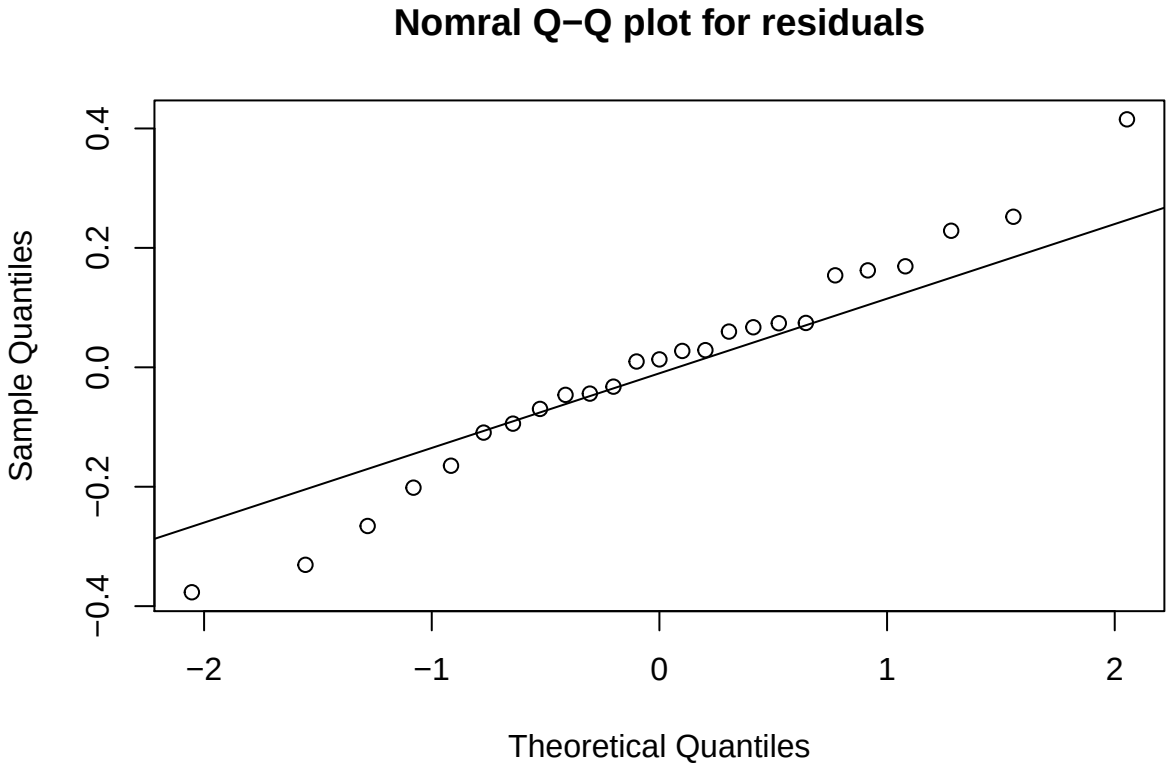

```
#legend(locator(n=1), legend=c("loess"), lty = 1)
```

Normality

```
boxplot(ep.hat1, main = "Boxplot: Residuals")
```




```
qqnorm(ep.hat1, main = "Nomral Q-Q plot for residuals")
qqline(ep.hat1)
```



(4) **Test Significant Relationship**

```
summary(lm(log.level~BloodLevel$Age))
```

```
##
## Call:
## lm(formula = log.level ~ BloodLevel$Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37670 -0.09435  0.01326  0.07428  0.41525
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.58498    0.06494  39.804  < 2e-16 ***
## BloodLevel$Age -0.22740    0.02651  -8.577 1.27e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1875 on 23 degrees of freedom
## Multiple R-squared:  0.7618, Adjusted R-squared:  0.7515
## F-statistic: 73.56 on 1 and 23 DF,  p-value: 1.273e-08
```

Null Hypothesis: beta1 = 0; Alternative Hypothesis: beta1 != 0, alpha = 0.05 Again, the relationship is significant because from the data we have, we can calculate the critical value of this T distribution. The estimated beta1 in this case is -0.22740 and SE of estimated beta1 is 0.02651, t* = (-0.22740-0)/0.02651 = -8.577, and since the value is less than -t(1-alpha/2, df) again (2.069), we reject the null hypothesis and conclude there is a statistically significant linear relationship between the two variables.

e) From the first set of graphs, we see that the lowess line is not as curved as the last part, also the lowess curve aligned more with the linera model. Moreoever, we can tell that now the residuals are more

normally distributed from the residual plot. Lastly, R^2 for the second analysis is about 75%, which is higher than the first analysis, which is about 66%.