HW5

Tianqi Wu

4/1/2020

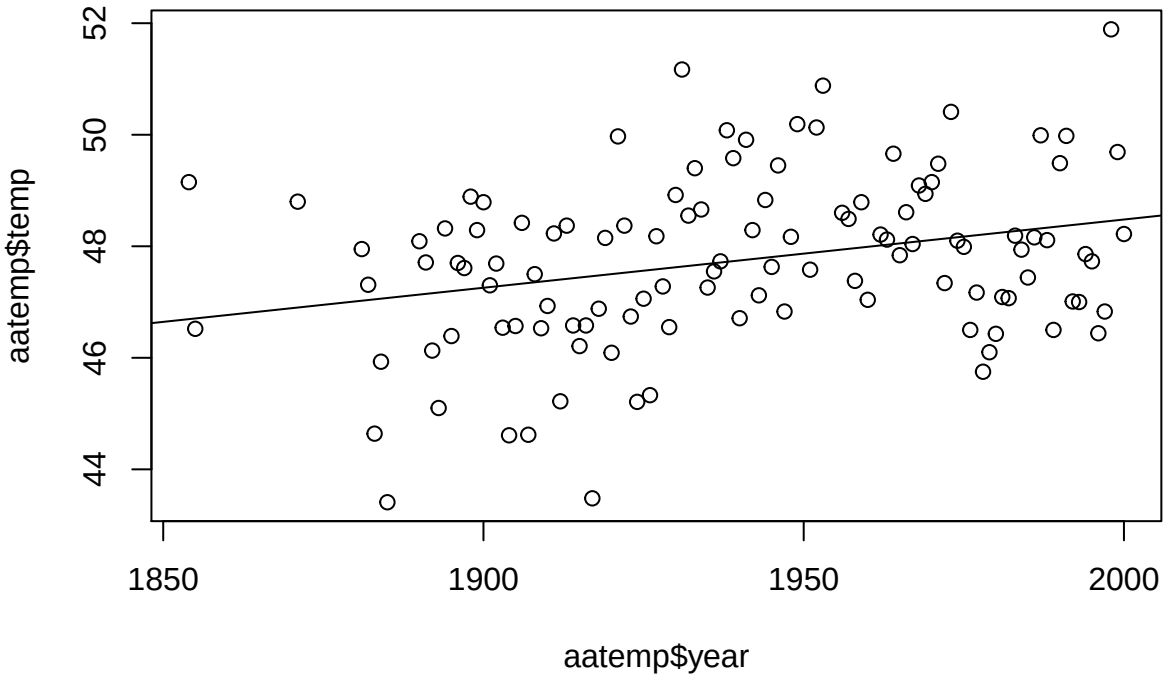
```
library(faraway)
library(nlme)
library(lmtest)
library(splines)
library(lmtest)
library(car)
attach(aatemp)
attach(infmort)
attach(pulp)
attach(chickwts)
```

Problem 1

1(a)

From the plot, there seems to be a weak positive linear trend between temperature and year.

```
plot(aatemp$year, aatemp$temp)
model.1a = lm(temp~year, data=aatemp)
abline(model.1a)
```



1(b)

Since p-value of D-W test is 0.01524 $<$ 0.05, it indicates that the errors are significantly correlated. After fitting a Regression with autocorrelated errors, the RSE is 1.475718 and it indicates that the model fits the data pretty well and it indicates there is a linear trend.

```
dwtest(model.1a)
```

```
##
##  Durbin-Watson test
##
## data:  model.1a
## DW = 1.6177, p-value = 0.01524
## alternative hypothesis: true autocorrelation is greater than 0
```

```
model.1b = gls(temp ~ year, correlation = corAR1(form= ~ year), data=aatemp)
summary(model.1b)
```

```
## Generalized least squares fit by REML
##   Model: temp ~ year
##   Data: aatemp
##        AIC     BIC    logLik
##   426.5694 437.479 -209.2847
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year
##  Parameter estimate(s):
##      Phi1
## 0.2303887
##
## Coefficients:
##                Value Std.Error  t-value p-value
```

```
## (Intercept) 25.18407  8.971864 2.807006  0.0059
## year         0.01164  0.004626 2.516015  0.0133
##
##  Correlation:
##      (Intr)
## year -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.7230803 -0.6321970 -0.0520135  0.6645795  2.3775123
##
## Residual standard error: 1.475718
## Degrees of freedom: 115 total; 113 residual
```

1(c)

Using backward elimination, degree is chosen to be 5. The temperature in 2020 is predicted to be 60.07774.

```
round(summary(lm(temp~poly(year, 10), aatemp))$coef[11,], dig=3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##      0.347      1.415      0.246      0.807
```

```
round(summary(lm(temp~poly(year, 9), aatemp))$coef[10,], dig=3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##      1.399      1.408      0.994      0.323
```

```
round(summary(lm(temp~poly(year, 8), aatemp))$coef[9,], dig=3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##     -1.101      1.408     -0.782      0.436
```

```
round(summary(lm(temp~poly(year, 7), aatemp))$coef[8,], dig=3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##     -0.937      1.406     -0.667      0.506
```

```
round(summary(lm(temp~poly(year, 6), aatemp))$coef[7,], dig=3)
```

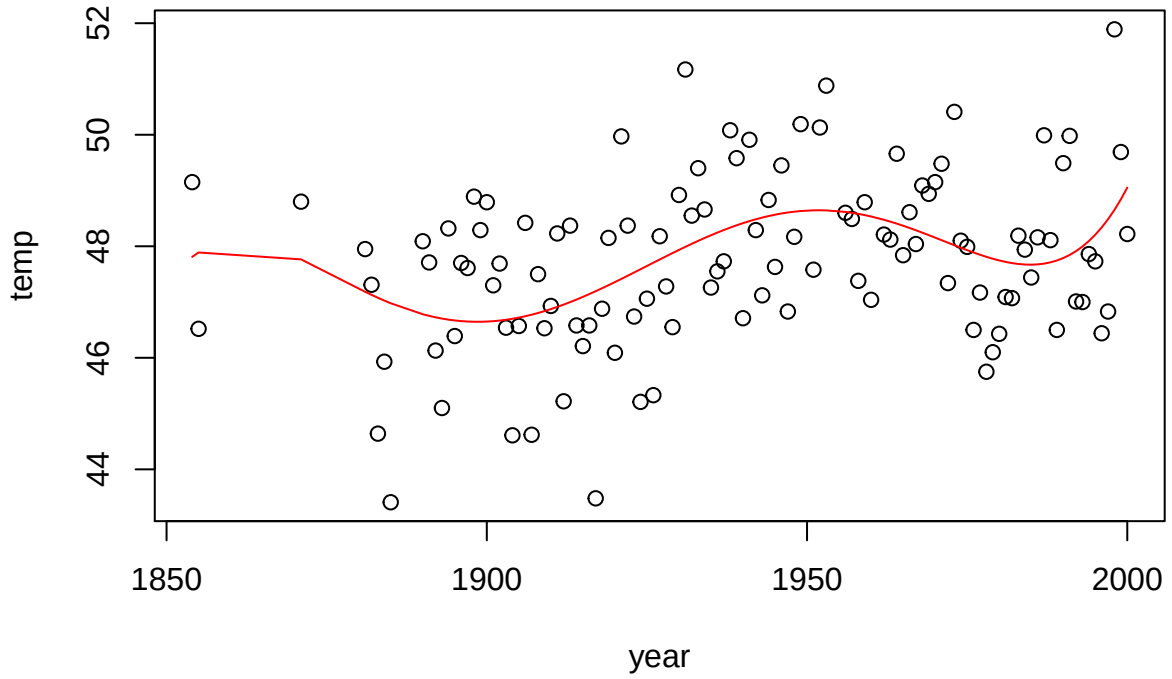
```
##   Estimate Std. Error    t value   Pr(>|t|)
##      1.212      1.402      0.865      0.389
```

```
round(summary(lm(temp~poly(year, 5), aatemp))$coef, dig=3)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      47.743      0.131 365.604    0.000
## poly(year, 5)1    4.762      1.400   3.400    0.001
## poly(year, 5)2   -0.907      1.400  -0.648    0.519
## poly(year, 5)3   -3.313      1.400  -2.366    0.020
## poly(year, 5)4    2.438      1.400   1.741    0.084
## poly(year, 5)5    3.382      1.400   2.415    0.017
```

```
model.1c = lm(temp~poly(year, 5), aatemp)

plot(temp~year,data=aatemp)
lines(aatemp$year, predict(model.1c),col="red", lty=1)
```



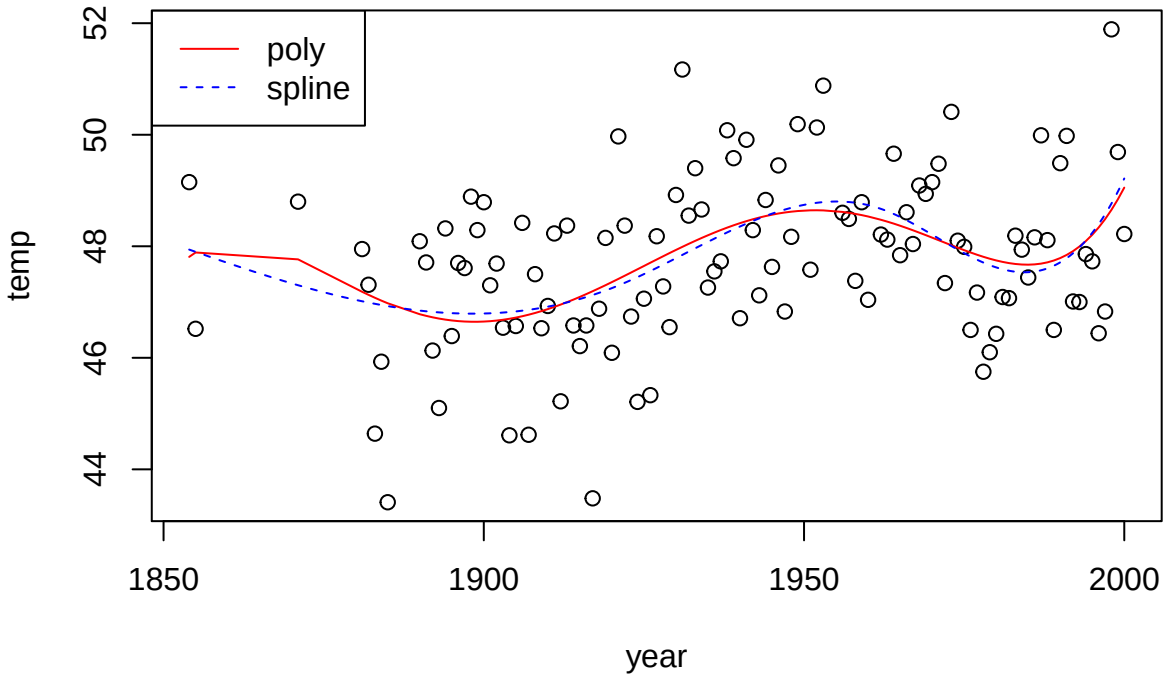

```
predict(model.1c, newdata=data.frame(year=2020))
```

```
##        1
## 60.07774
```

1(d)

There is not a big difference between the fitting of cubic spline model and polynomial model. It seems that cubic spline model is smoother and might be better in this case.

```
model.1d = lm(temp~bs(year, df=6,intercept=TRUE), data=aatemp)
plot(temp~year,data=aatemp)
lines(aatemp$year, predict(model.1c),col="red", lty=1, )
lines(spline(aatemp$year, predict(model.1d)), col="blue", lty=2)
legend("topleft", col=c("red", "blue"), lty=c(1,2), legend=c("poly", "spline"))
```

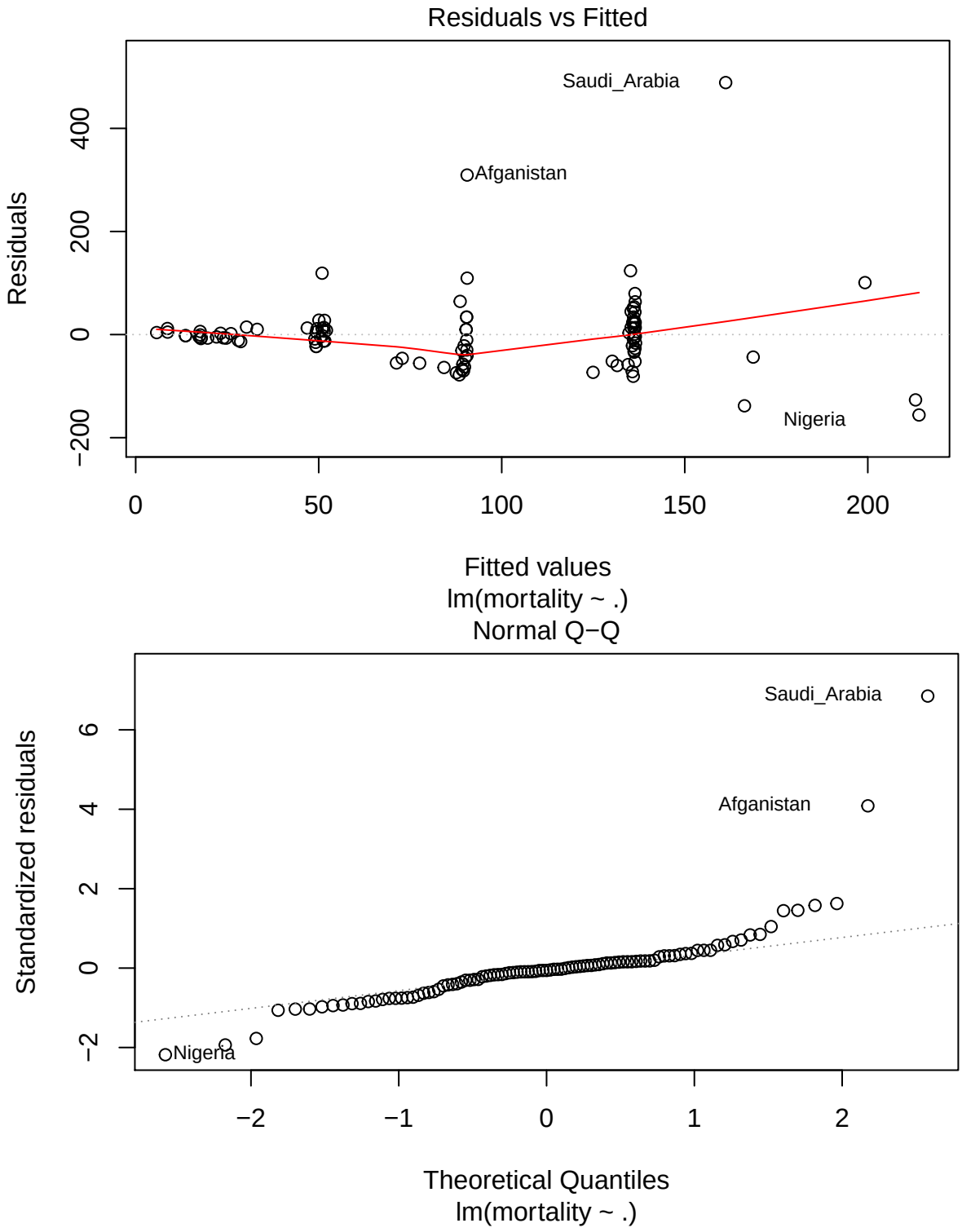


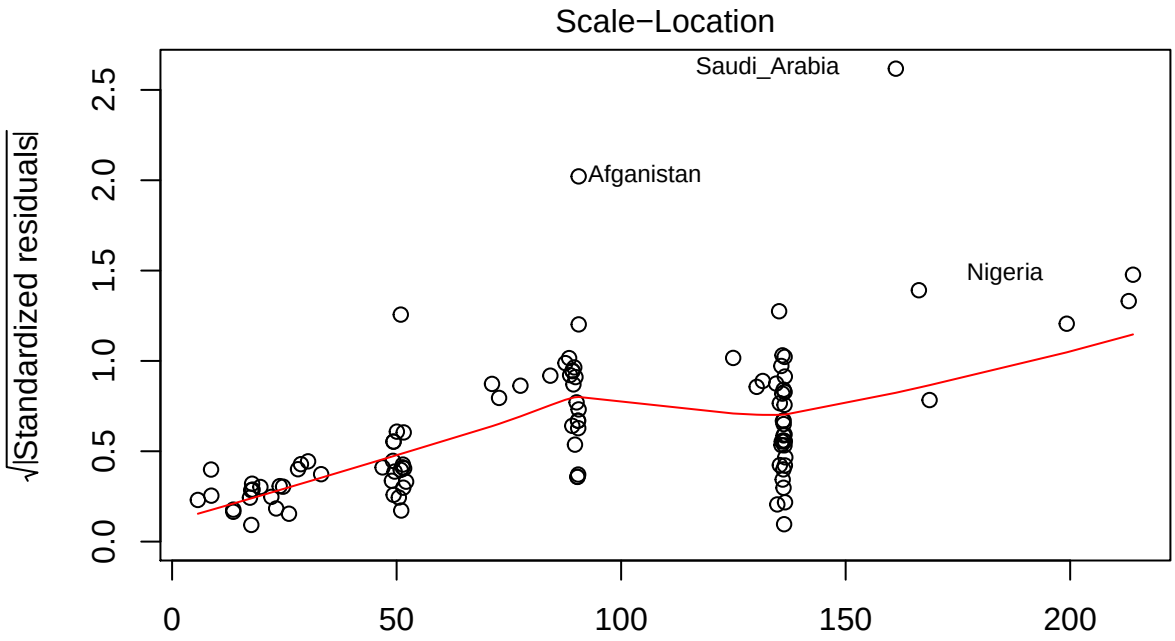
Problem 2

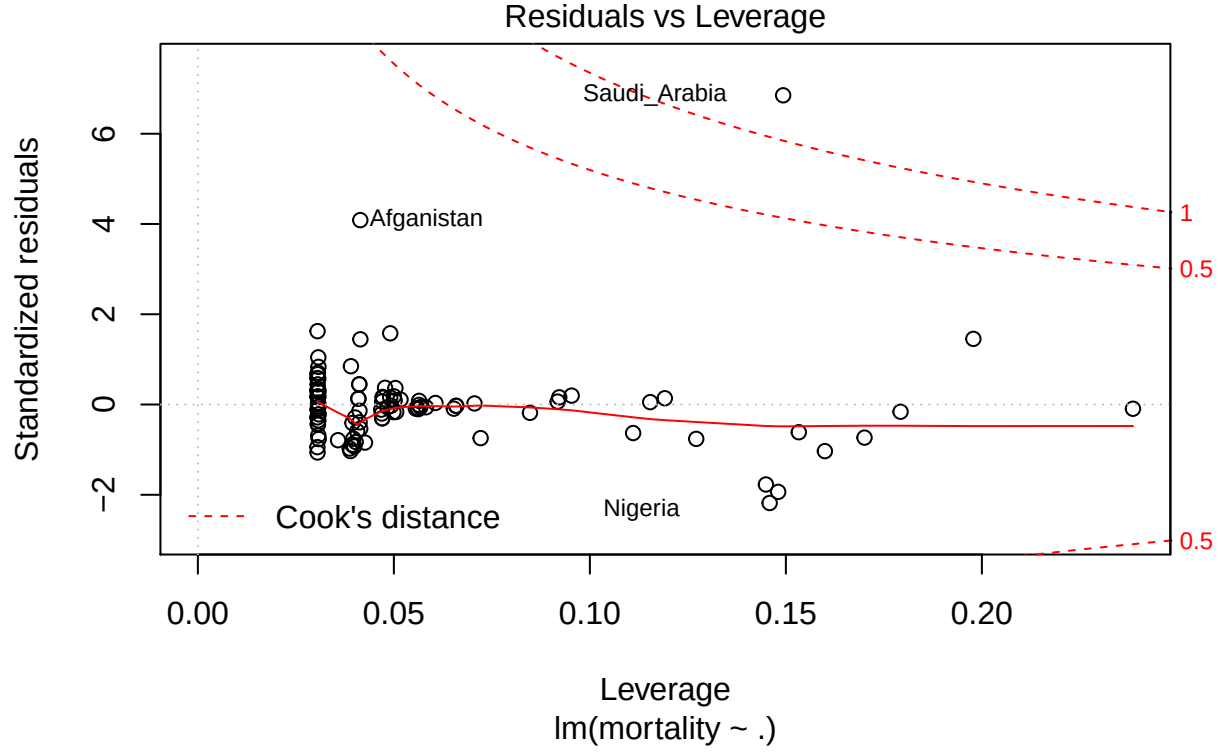
From the model diagnostic for lm(mortality ~ .,data=infmort), we can see that 'Saudi_Arabia' and 'Afganistan' deviate a lot from other samples and can be seen as unusual points. Hence, we remove them to fit a better model. There is no obvious linear trend between motarlity and income in the original data but there is an obvious linear trend if we perform the log transformation. After log transformation and removing unusual points, we refit the model and the adjusted R-squared is 0.7124.

Infant mortality is expected to be 6.9317 if the region is Africa and the country exports oil without considering income. In addition, infant mortality is expected to decrease by 0.3483 with one unit increase of log(income). Infant mortality is expected to decrease by 1.0396, 0.8749 and 0.5376 if the region is Europe, Asia and Americas respectively. Infant mortality is expected to decrease by 0.3070 if the country does not export oil.

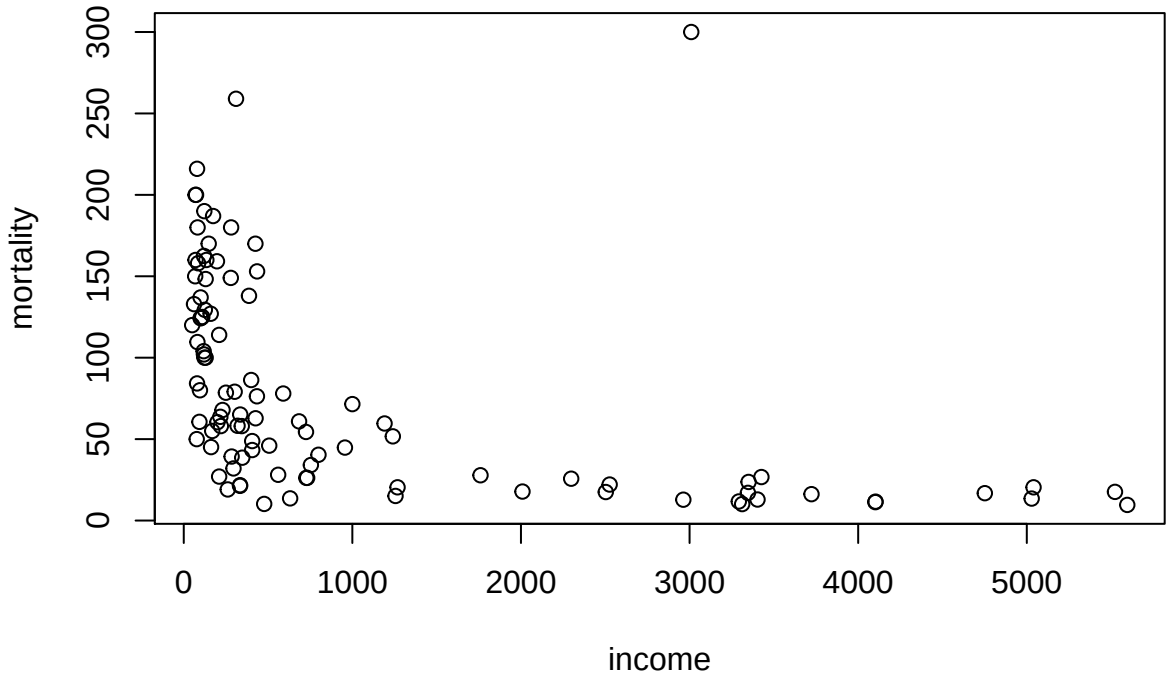
```
model.2a = lm(mortality ~ .,data=infmort)
plot(model.2a)
```


Residuals vs Fitted
Saudi_Arabia
Afganistan
Nigeria
Residuals
400
200
0
−200
0
50
100
150
200
Fitted values
lm(mortality ~ .)
Normal Q−Q
Saudi_Arabia
Afganistan
Nigeria
Standardized residuals
6
4
2
0
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(mortality ~ .)

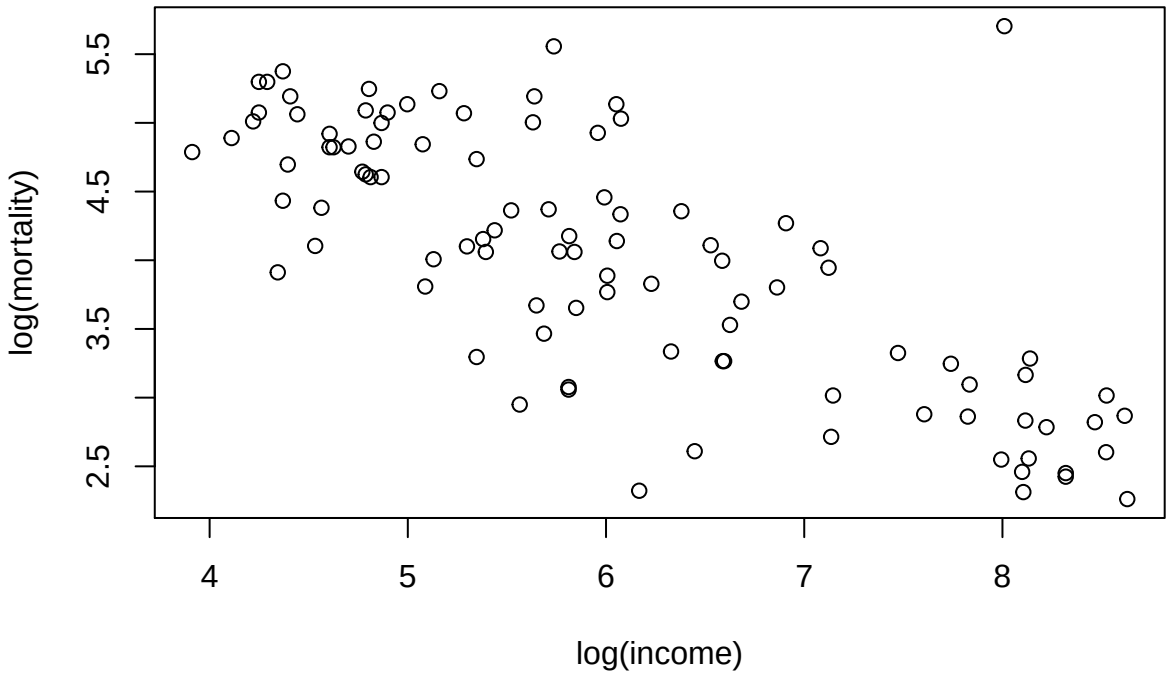





```
data.2 = infmort[!(row.names(infmort) %in% c('Saudi_Arabia       ', 'Afganistan         ')),]
plot(mortality~income,data=data.2)
```




```
plot(log(mortality) ~ log(income),data=data.2)
```




```
model.2b = lm(log(mortality) ~ log(income)+region+oil, data=data.2)
summary(model.2b)
```

```
##
## Call:
## lm(formula = log(mortality) ~ log(income) + region + oil, data = data.2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.27903 -0.31040 -0.02586  0.29594  1.56221
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.9317     0.3755  18.460  < 2e-16 ***
## log(income)        -0.3483     0.0570  -6.112 2.28e-08 ***
## regionEurope       -1.0396     0.2174  -4.782 6.50e-06 ***
## regionAsia         -0.8749     0.1354  -6.461 4.73e-09 ***
## regionAmericas     -0.5376     0.1557  -3.452 0.000839 ***
## oilno oil exports  -0.3070     0.1989  -1.543 0.126203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 93 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7271, Adjusted R-squared:  0.7124
## F-statistic: 49.55 on 5 and 93 DF,  p-value: < 2.2e-16
```

Problem 3

From the anova result, p-value = 0.0226 < 0.05 indicates that there is difference between the operators. Operator d has p-value 0.0486 < 0.05 and it is statistically significant to the brightness.

The brightness is expected to be 60.2400 if the pulp is operator a. In addition, the brightness is expected to be decreased by 0.18 if the pulp is operator b and increased by 0.38 and 0.44 if the pulp is operator c and d respectively. It means that pulp with operator d is expected to have the largest brightness and pulp with operator d is expected to have the lowest brightest.

```
model.3 = lm(bright~operator, data=pulp)
anova(model.3)
```

```
## Analysis of Variance Table
##
## Response: bright
##           Df Sum Sq Mean Sq F value  Pr(>F)
## operator   3   1.34 0.44667  4.2039 0.02261 *
## Residuals 16   1.70 0.10625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model.3)
```

```
##
## Call:
## lm(formula = bright ~ operator, data = pulp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.440 -0.195 -0.070  0.175  0.560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.2400     0.1458 413.243   <2e-16 ***
## operatorb    -0.1800     0.2062  -0.873   0.3955
```

```
## operatorc     0.3800     0.2062   1.843   0.0839 .
## operatord     0.4400     0.2062   2.134   0.0486 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.326 on 16 degrees of freedom
## Multiple R-squared:  0.4408, Adjusted R-squared:  0.3359
## F-statistic: 4.204 on 3 and 16 DF,  p-value: 0.02261
```

Problem 4

From the anova result, p-value = 5.936e-10 < 0.05 indicates that there is difference between the feed type.

Since the p-value for BP test is 0.4958>0.05, we fail to reject the null hypothesis: H0: There is homocedasticity. The constant variance assumption is not violated.

The normal qqplot shows that the residuals are not departing from the normality assumption in the central part of the data distribution. There are some deviations in the left tail and right tail. We failed to reject the wilks-shapiro test; therefore we support the null hypothesis of normality of the response variable. There is no obvious unusual points with high leverage or cook's distance.

```
model.4 = lm(weight~feed, data=chickwts)
anova(model.4)
```

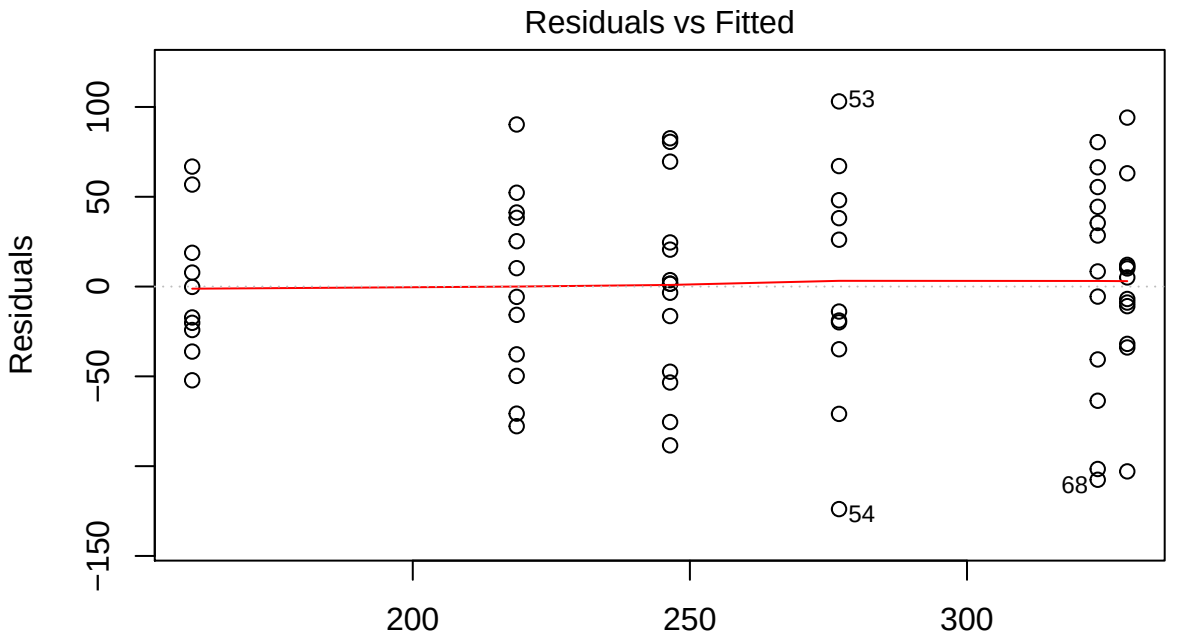
```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## feed       5 231129   46226  15.365 5.936e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

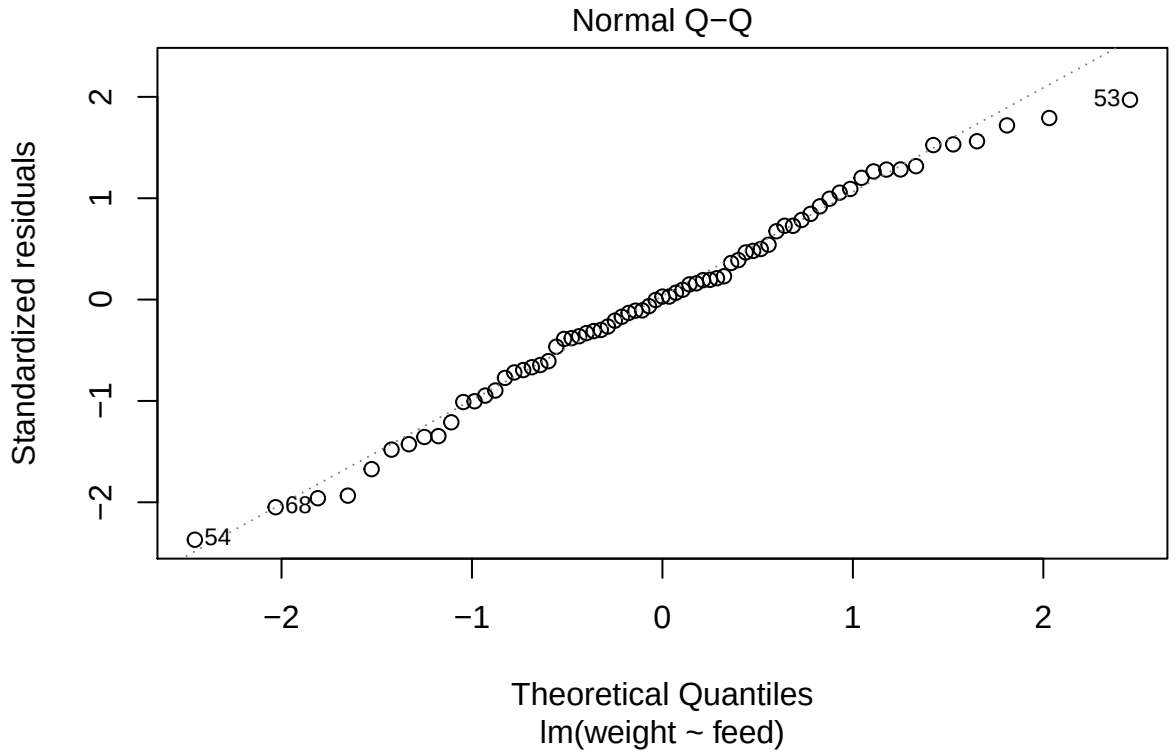
```
summary(model.4)
```

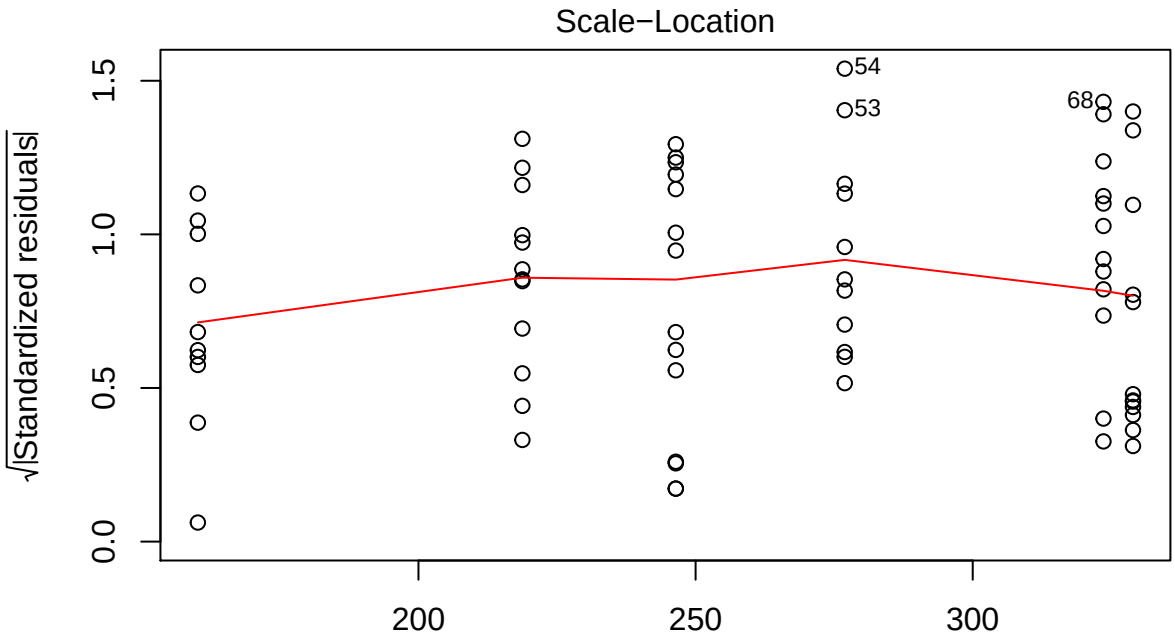
```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    323.583     15.834  20.436  < 2e-16 ***
## feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
## feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
## feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
## feedsoybean    -77.155     21.578  -3.576 0.000665 ***
## feedsunflower    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
```

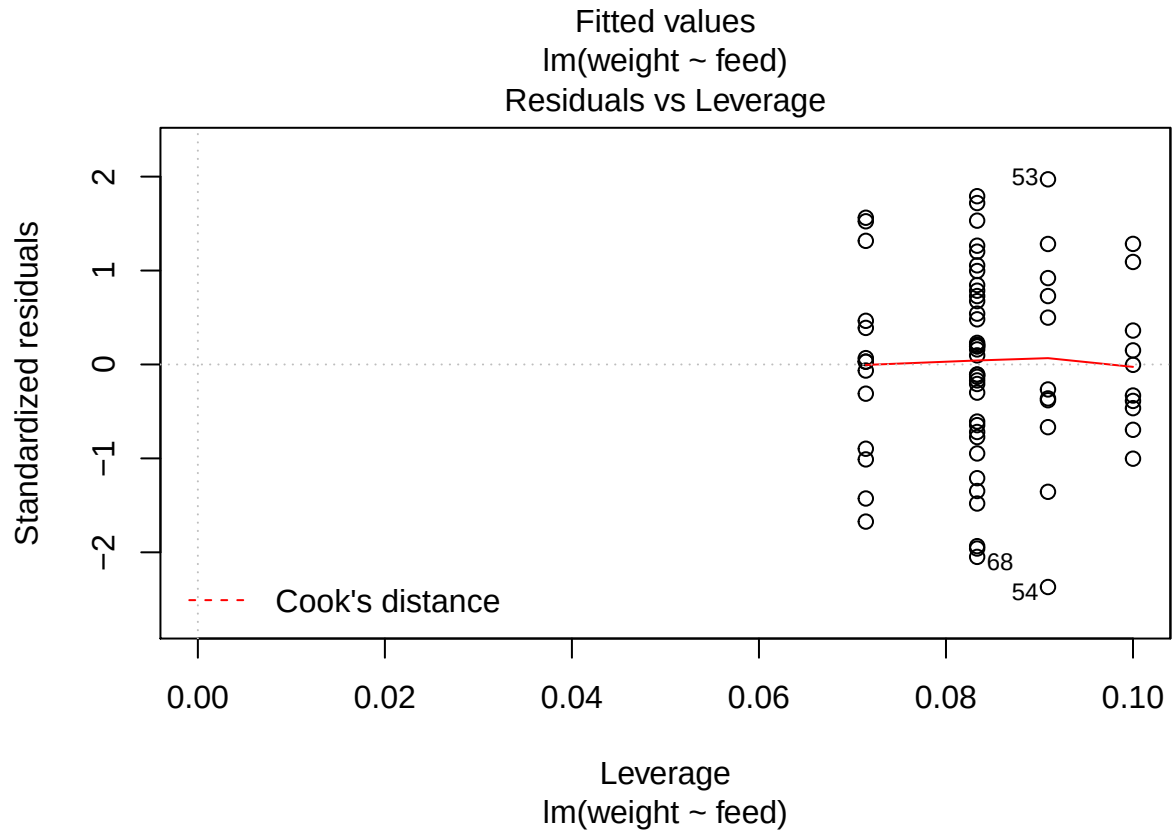
```
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

```
plot(model.4)
```










```
# constant-variance
bptest(model.4)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  model.4
## BP = 4.3822, df = 5, p-value = 0.4958
```

```
# normality
shapiro.test(residuals(model.4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.4)
## W = 0.98616, p-value = 0.6272
```