MIdterm2

Tianqi Wu

4/8/2020

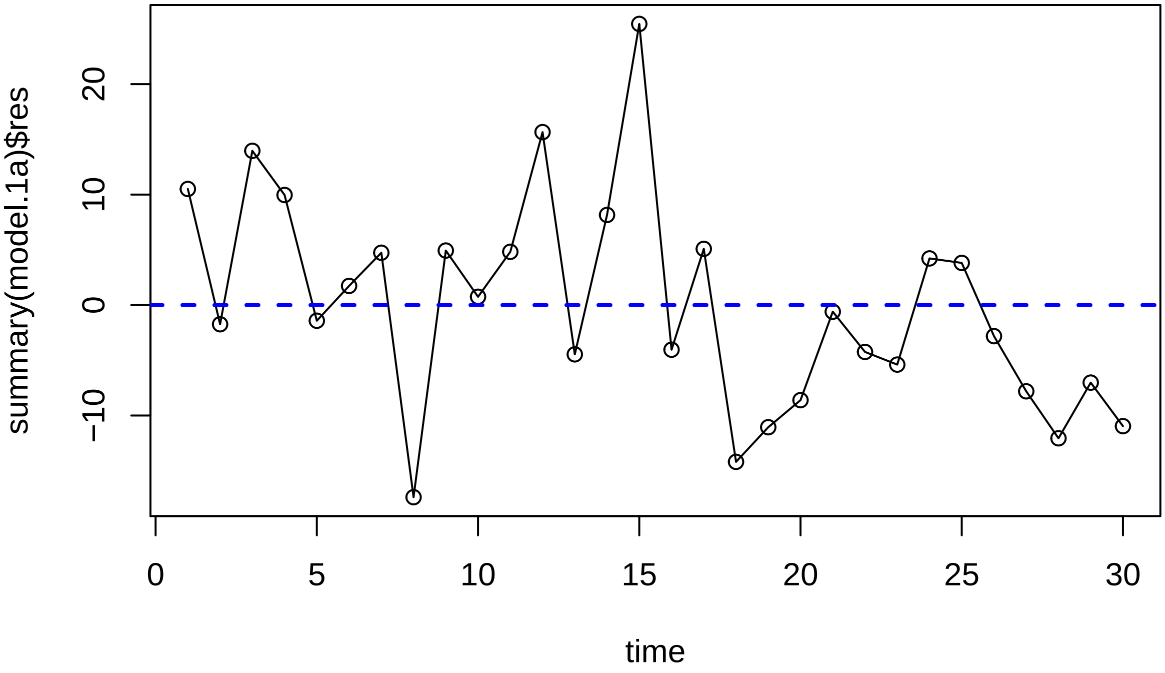
```
library(faraway)
library(lmtest)
library(nlme)
library(MASS)
attach(cheddar)
attach(cars)
```

Problem 1

1(a)

There is no obvious pattern between residuals of the model and time. But it seems that taste score decreases as time increases. There might be some trend and errors might be correlated.

```
model.1a = lm(taste~., data=cheddar)
cheddar$time = 1:nrow(cheddar)
plot(summary(model.1a)$res~time, type='o', data=cheddar)
abline(h=0, lty=2, col="blue", lwd=2)
```



1(b)

For dwtest, p-value > 0.05 indicates suggests that errors are not correlated. The RSS for the gls model is 10.33276 and it indicates that the model does not fit data well. The CI for phi includes the zero and it indicates that autocorrelation may not be needed.

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

```
##
##  Durbin-Watson test
##
## data:  model.1a
## DW = 1.5751, p-value = 0.08869
## alternative hypothesis: true autocorrelation is greater than 0
```

```
model.1b = gls(taste~.-time, corAR1(form= ~ time), data=cheddar)
summary(model.1b)
```

```
## Generalized least squares fit by REML
##   Model: taste ~ . - time
##   Data: cheddar
##      AIC      BIC  logLik
##   214.94 222.4886 -101.47
##
## Correlation Structure: AR(1)
##  Formula: ~time
##  Parameter estimate(s):
##       Phi
## 0.2641944
##
## Coefficients:
##                  Value Std.Error   t-value p-value
## (Intercept) -30.332472 20.273077 -1.496195  0.1466
## Acetic        1.436411  4.876581  0.294553  0.7707
## H2S           4.058880  1.314283  3.088284  0.0047
## Lactic       15.826468  9.235404  1.713674  0.0985
##
##  Correlation:
##        (Intr) Acetic H2S
## Acetic -0.899
## H2S     0.424 -0.395
## Lactic  0.063 -0.416 -0.435
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -1.64546468 -0.63861716 -0.06641714  0.52255676  2.41323021
##
## Residual standard error: 10.33276
## Degrees of freedom: 30 total; 26 residual
```

```
intervals(model.1b,which = "var-cov")
```

```
## Approximate 95% confidence intervals
##
##  Correlation structure:
##          lower      est.     upper
## Phi -0.1690265 0.2641944 0.6118599
## attr(,"label")
## [1] "Correlation structure:"
##
##  Residual standard error:
##    lower     est.    upper
##  7.62646 10.33276 13.99940
```

1(c)

If we fit a LS model but with time now as an additional predictor, the predictor time has p-value < 0.05 indicates that it is statistically significant at 5% level.

```
model.1c = lm(taste~., data=cheddar)
summary(model.1c)
```

```
##
## Call:
## lm(formula = taste ~ ., data = cheddar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3523  -4.9735  -0.5089   4.8531  23.1311
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.6127    17.9845  -2.036  0.05250 .
## Acetic        4.1275     4.2556   0.970  0.34139
## H2S           3.5387     1.1315   3.127  0.00444 **
## Lactic       17.9527     7.7875   2.305  0.02973 *
## time         -0.5459     0.2043  -2.672  0.01306 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.112 on 25 degrees of freedom
## Multiple R-squared:  0.7291, Adjusted R-squared:  0.6858
## F-statistic: 16.83 on 4 and 25 DF,  p-value: 8.205e-07
```

1(d)

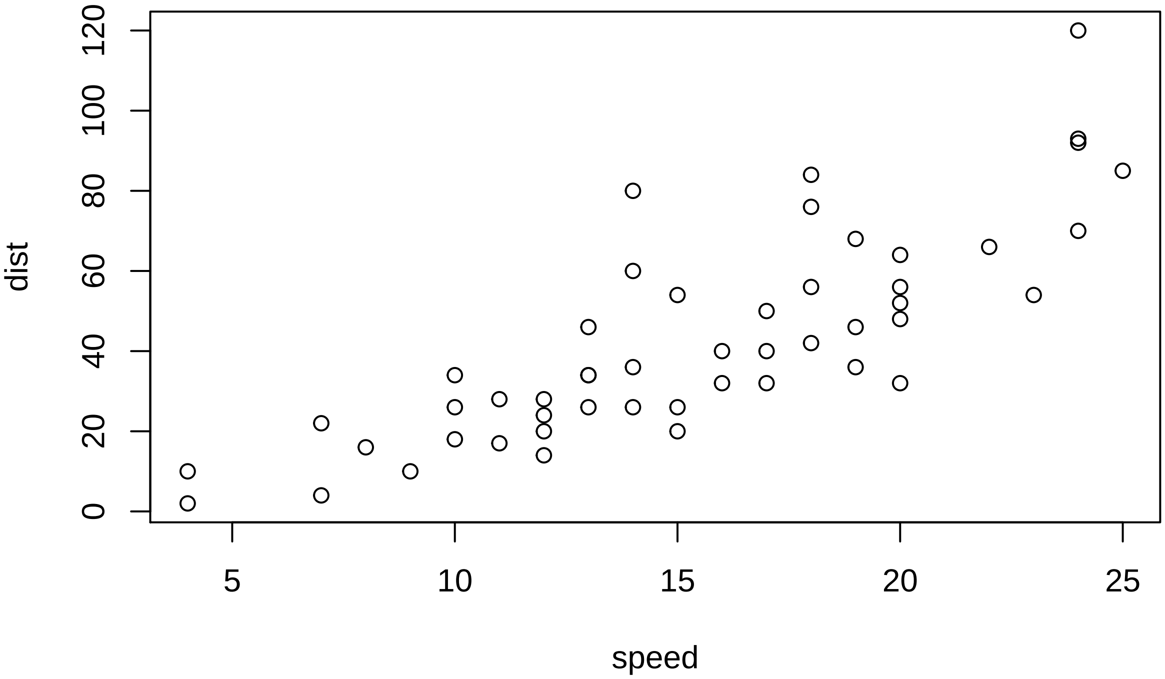
For LS model, we assume errors are independent with constant variance and use time as a predictor explicitly. Taste is expected to be decreased by 0.5459 if time increases by 1.

For GLS model, we assume errors are correlated and the covariance matrix takes some particular form. The pattern of time is included in the correlation structure. In this case, we use AR(1) time series.

Problem 2

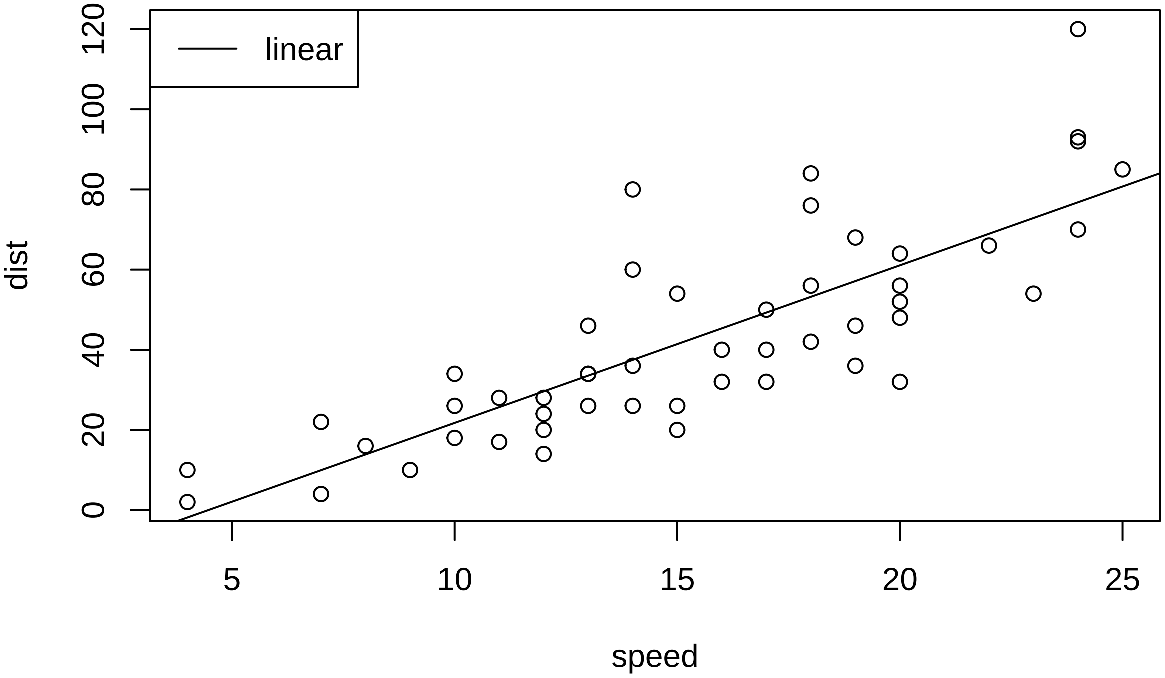
2(a)

```
plot(dist~speed, data=cars)
```



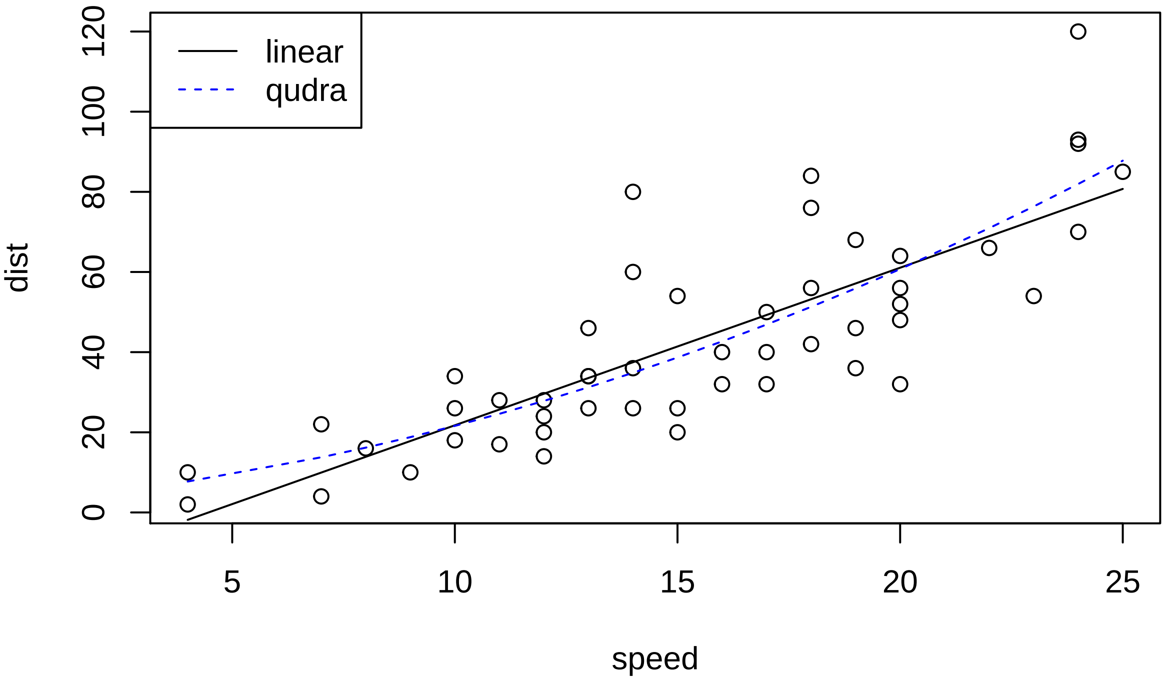
2(b)

```
model.2b = lm(dist~speed, data=cars)
plot(dist~speed, data=cars)
abline(model.2b)
legend("topleft", col=c("black"), lty=c(1), legend=c("linear"))
```



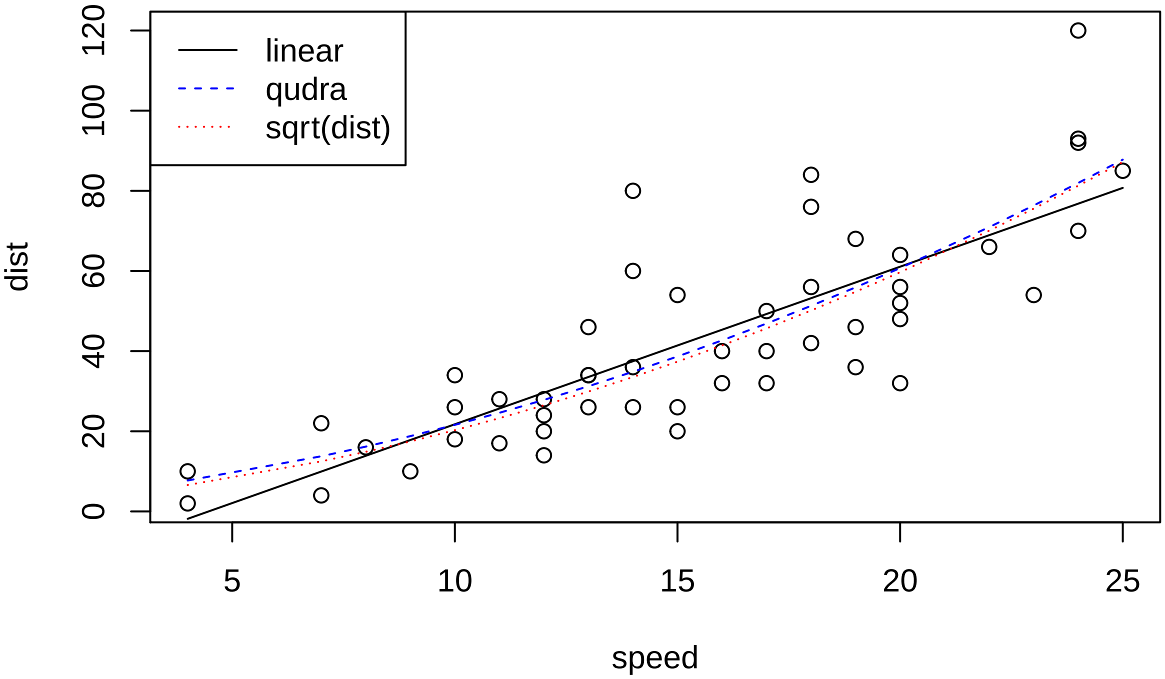
2(c)

```r
model.2c = lm(dist~speed+I(speed^2), data=cars)
plot(dist~speed, data=cars)
lines(cars$speed, predict(model.2b), col="black", lty=1)
lines(cars$speed, predict(model.2c), col="blue", lty=2)
legend("topleft", col=c("black", "blue"), lty=c(1,2), legend=c("linear", "qudra"))
```



2(d)

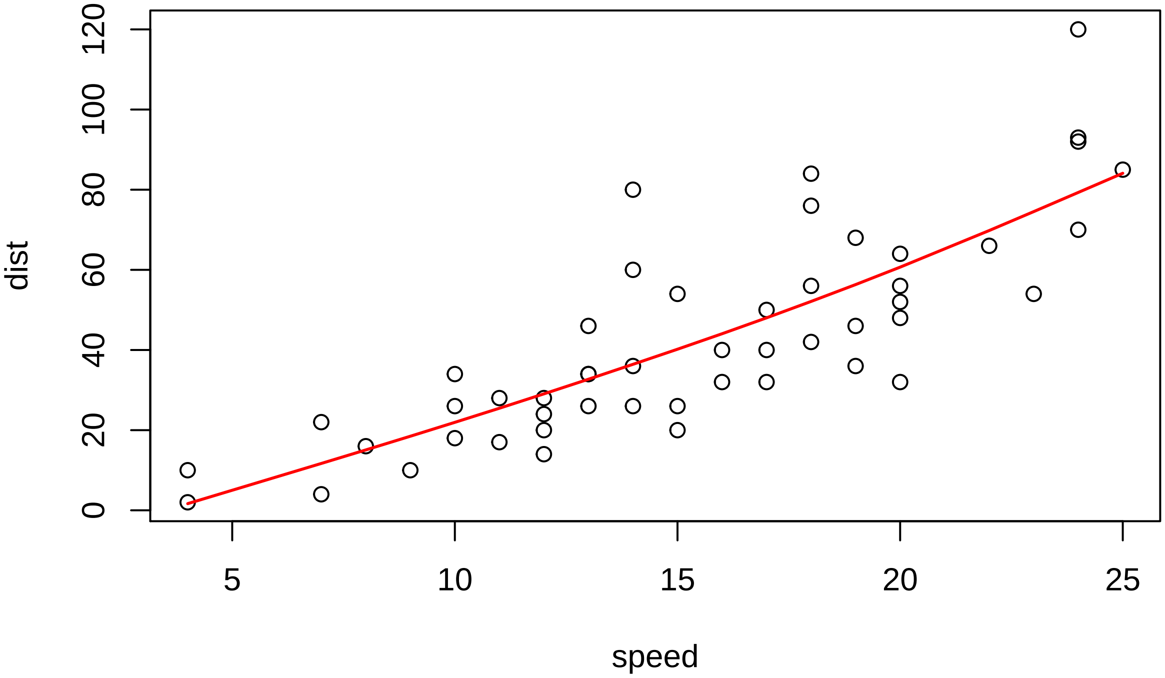
```r
model.2d = lm(sqrt(dist)~speed, data=cars)
plot(dist~speed, data=cars)
lines(cars$speed, predict(model.2b), col="black", lty=1)
lines(cars$speed, predict(model.2c), col="blue", lty=2)
lines(cars$speed, predict(model.2d)^2, col="red", lty=3)
legend("topleft", col=c("black", "blue",'red'), lty=c(1,2,3), legend=c("linear", "qudra",'sqrt(dist)'))
```



2(e)

The default smoothing spline fit is similar to quadratic fit and sqrt(dist) as response fit. It seems that smoothing spline fits a little better at boundaries.

```
plot(dist~speed, data=cars)
lines(smooth.spline(cars$speed,cars$dist), lwd=1.5, col="red")
```

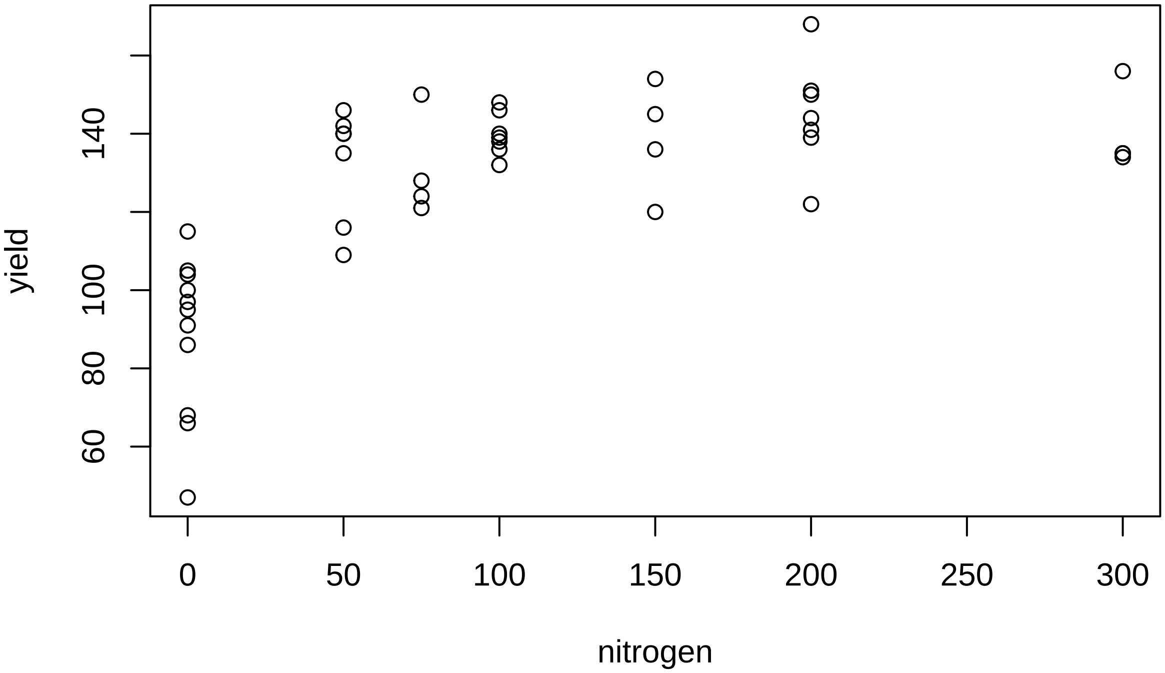


Problem 3

We fist plot yield against nitrogen and we find that there seems to be a linear trend but we need transformation to check. The data has replicates and it indicates that lack-of-fit test is suitable. The p-value for Shapiro-Wilk test and Breusch-Pagan test are greater than 0.05 and it means that the normality and homocedasticity assumptions are not violated.

Then, we plot yield against log(nitrogen+1) and the linear trend is much more obvious now. We use the transformation and refit the model. The residual plot shows normal behavior. The adjusted R-squared increases from 0.3818 to 0.6985.Finally, we perform a lack-of-fit test and the p-value is 0.843 > 0.05 and it means that we fail to reject the null hypothesis and concludes that our model does not have lack-of-fit. Hence, log transformation is suitable for this problem.

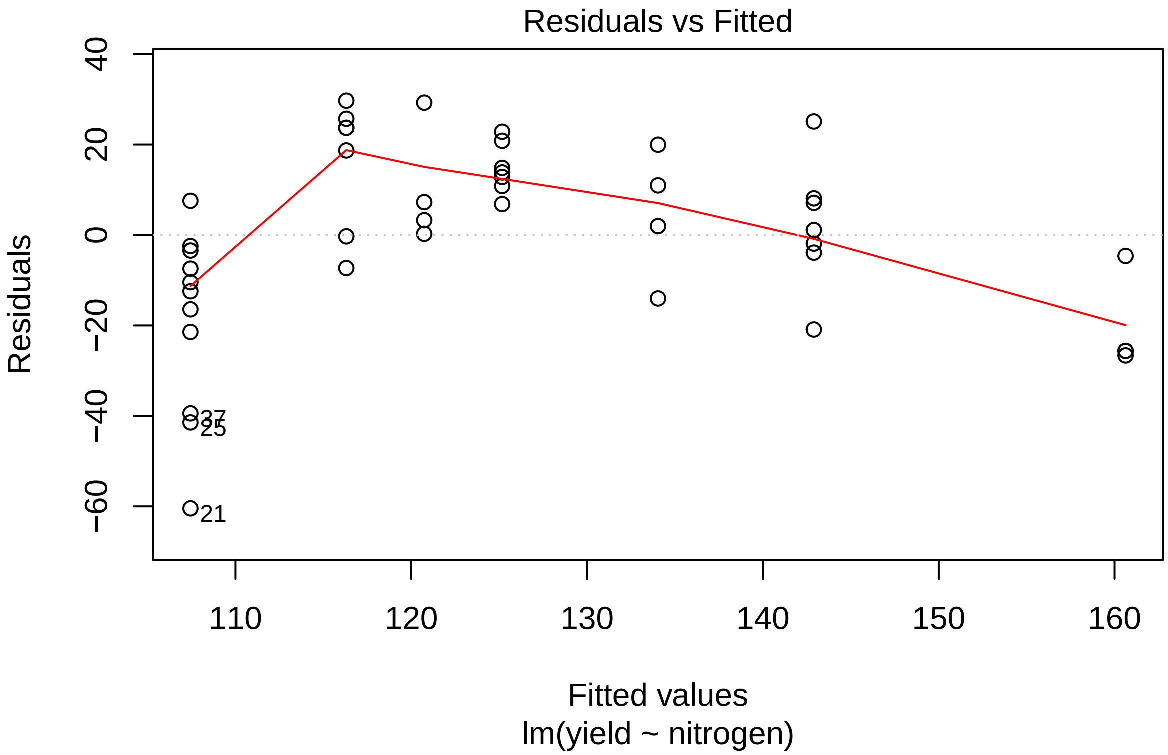
```
plot(yield~nitrogen, data=cornnit)
```



```
model.3.lm = lm(yield~nitrogen, data=cornnit)
summary(model.3.lm)
```

```
##
## Call:
## lm(formula = yield ~ nitrogen, data = cornnit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.439 -10.939   1.534  14.082  29.697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.43864    4.66622   23.02  < 2e-16 ***
## nitrogen      0.17730    0.03377    5.25 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.53 on 42 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3818
## F-statistic: 27.56 on 1 and 42 DF,  p-value: 4.713e-06
```

```
plot(model.3.lm,1)
```



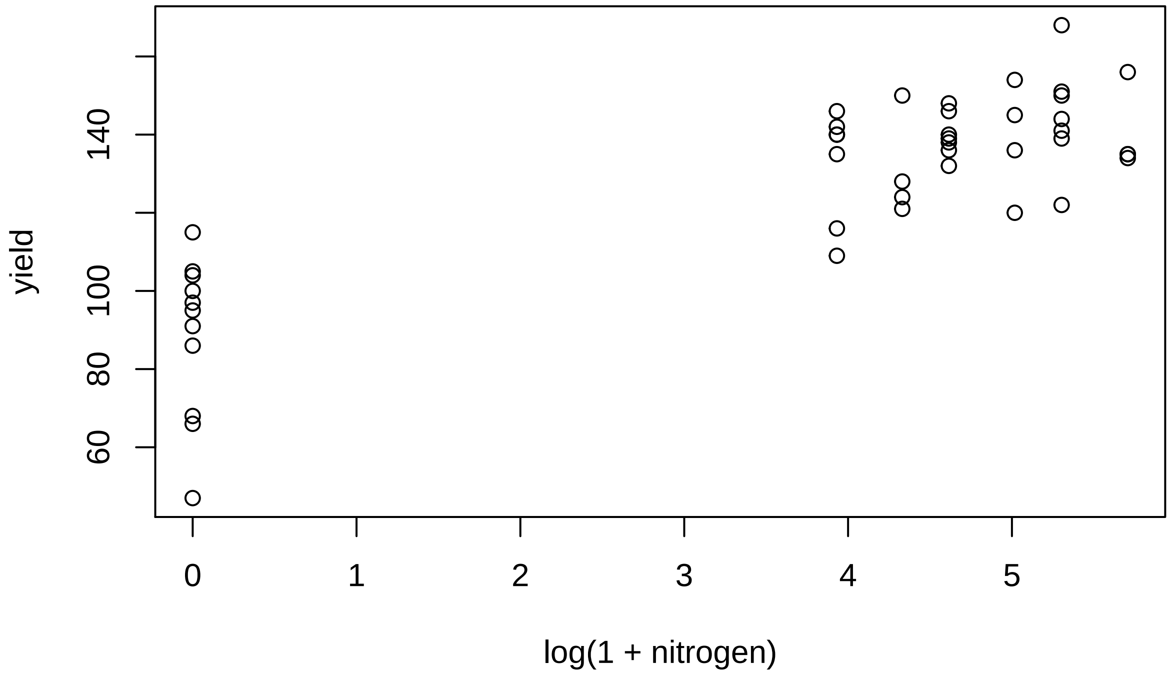
```
## Normality test
shapiro.test(residuals(model.3.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.3.lm)
## W = 0.95164, p-value = 0.06332
```

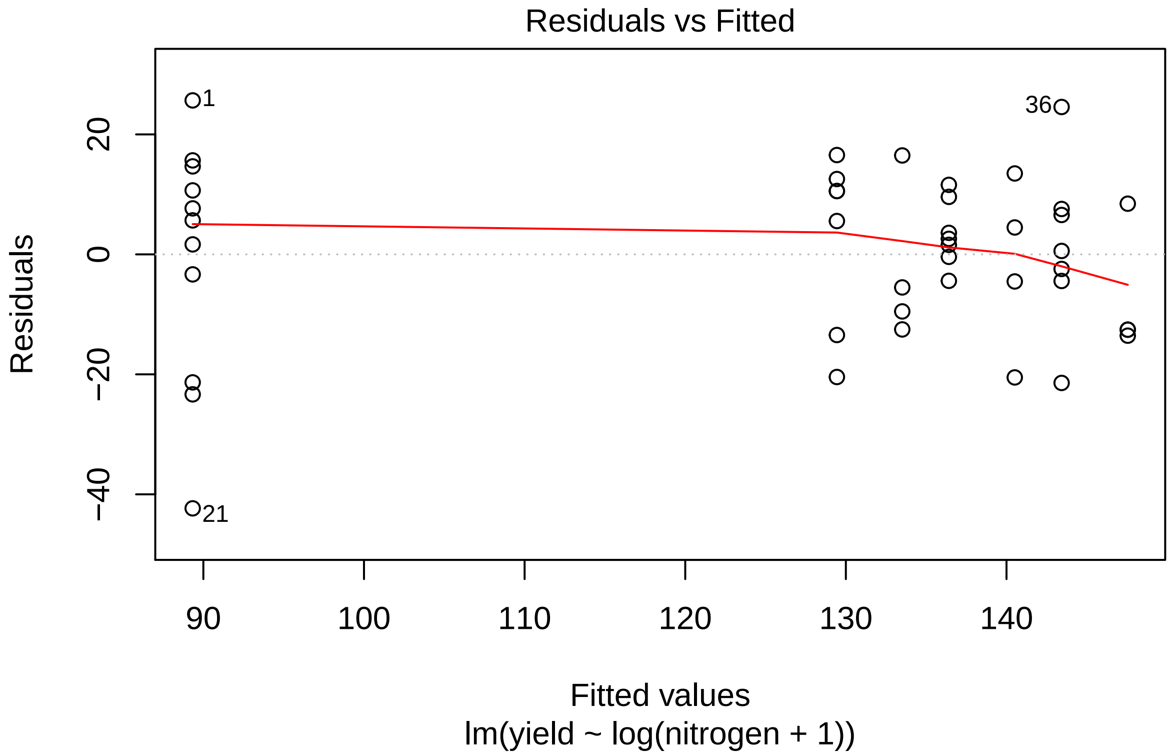
```
## homocedasticity test
bptest(model.3.lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model.3.lm
## BP = 1.5558, df = 1, p-value = 0.2123
```

```
## log tranformation fit
plot(yield~log(1+nitrogen), data=cornnit)
```



```
model.3.log = lm(yield~log(nitrogen+1), data=cornnit)
plot(model.3.log,1)
```



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

```
##
## Call:
## lm(formula = yield ~ log(nitrogen + 1), data = cornnit)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -42.335 -10.261   2.126  10.558  25.665 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         89.335      4.227   21.13  < 2e-16 ***
## log(nitrogen + 1)   10.201      1.017   10.03 1.03e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14.34 on 42 degrees of freedom
## Multiple R-squared:  0.7055, Adjusted R-squared:  0.6985 
## F-statistic: 100.6 on 1 and 42 DF,  p-value: 1.025e-12
```

```
## Goodness-of-fit test
model3.factor = lm(yield~factor(nitrogen), data=cornnit)
anova(model.3.log,model3.factor)
```

```
## Analysis of Variance Table
## 
## Model 1: yield ~ log(nitrogen + 1)
## Model 2: yield ~ factor(nitrogen)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     42 8633.5                           
## 2     37 8186.8  5    446.72 0.4038  0.843
```