Math 327 Chapter 4 Homework

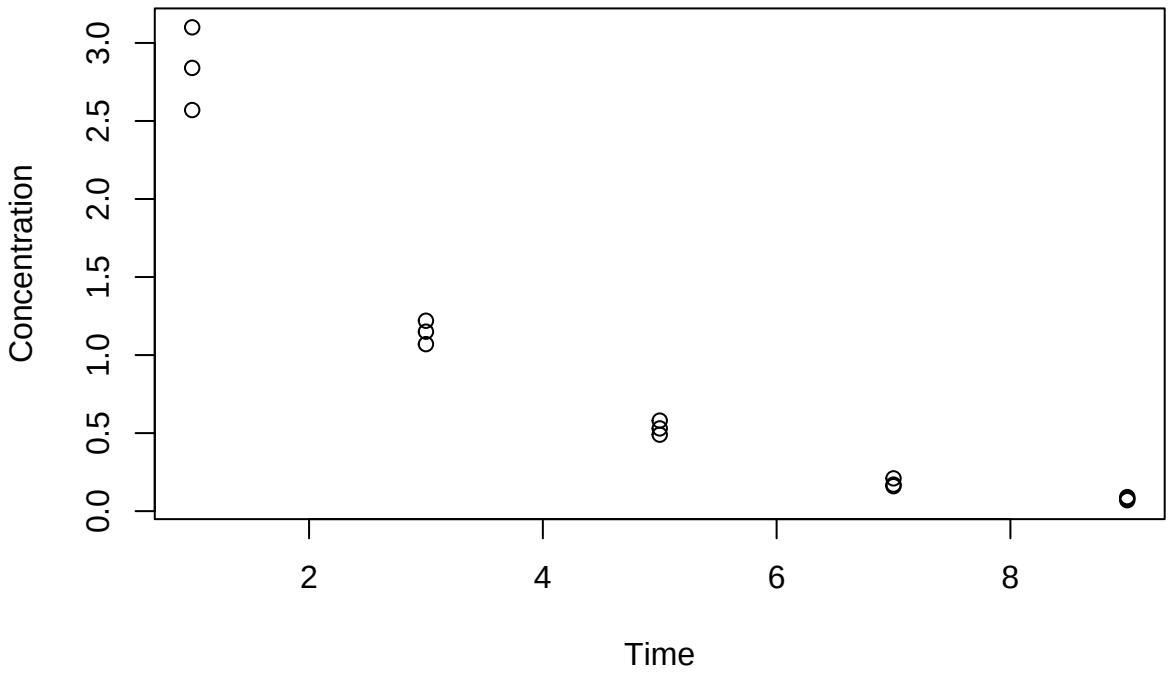
David Oniani

Name: David Oniani

Note: this data set is from the book, Applied Linear Statistical Models, but Kutner et al, Chapter 3, Problems 3.15-3.16. The data are from a designed biological experiment where the conentration of a drug (measured in ng/ml) was measured 3 times each at 1, 3, 5, 7, and 9 minutes.

In the first model, a simple linear regression model is used to fit Concentration vs Time:

```
# Open the data file, using the appropriate path for your computer
# The following load() function will create the data frame, hw4data,
# with columns, Concentration and Time.

load ("./hw4data.Rdata")

# Plot the data

plot (Concentration ~ Time, data=hw4data)
```




```
# Fit the linear model

cofit <- lm (Concentration ~ Time, data=hw4data)
summary (cofit)
```

```
##
## Call:
## lm(formula = Concentration ~ Time, data = hw4data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -0.5333 -0.4043 -0.1373  0.4157  0.8487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5753     0.2487  10.354 1.20e-07 ***
## Time         -0.3240     0.0433  -7.483 4.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4743 on 13 degrees of freedom
## Multiple R-squared:  0.8116, Adjusted R-squared:  0.7971
## F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06
```

```
# Added confidence intervals
# This is not a part of the original code
confint (cofit)
```

```
##                 2.5 %     97.5 %
## (Intercept)  2.0379803  3.1126864
## Time        -0.4175412 -0.2304588
```

Q1: Interpret the numeric results above.

The mean response value when Time (in minutes) is 0 equals 2.575 ng/ml and is between 2.038 ng/ml and 3.113 ng/ml with 95% confidence.

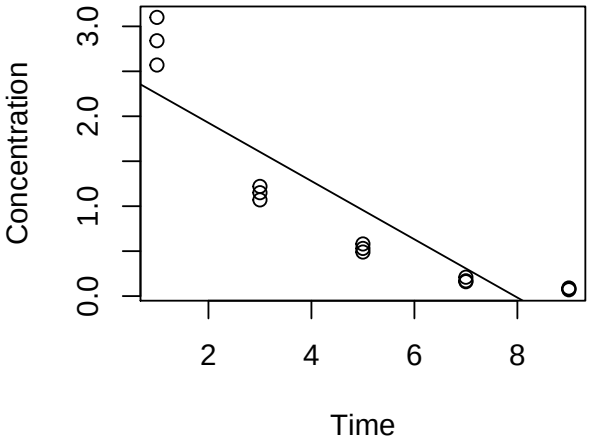
The estimated mean response value is -0.324 ng/ml per minutes which changes between -0.418 ng/ml per minutes and -0.230 ng/ml per minutes, for any 1-unit increase in the predictor, with 95% confidence.

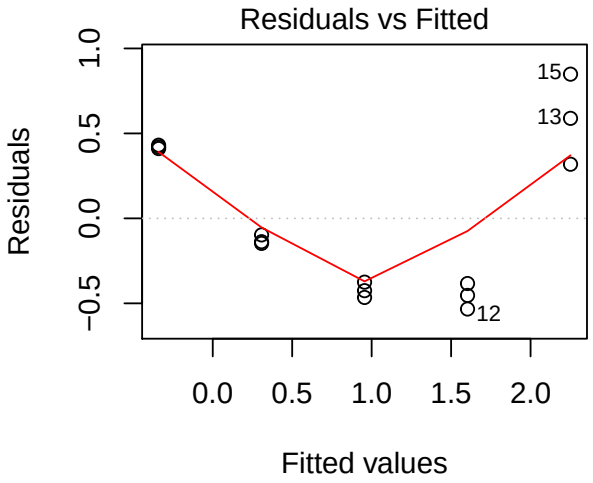
It should be noted that both intercept and slope are significant estimates with p-values of 1.20×10^{-7} and 4.61×10^{-6} respectively.

The Multiple R-squared value is 0.8116 which tells us that 81.16% of total variation in Concentration is explained by Time.

Residual standard error is 0.4743 meaning that on average, predictions of the model are 0.4743 ng/ml off from the actual value.

```
# The statement par (mfrow = c(1,2)) puts two plots side-by-side

par (mfrow = c(1,2))

# Plot the data and the fitted line
plot (Concentration ~ Time, data=hw4data)
abline (cofit)

# Plot the residuals vs the fitted values

plot (cofit, which=1)
```





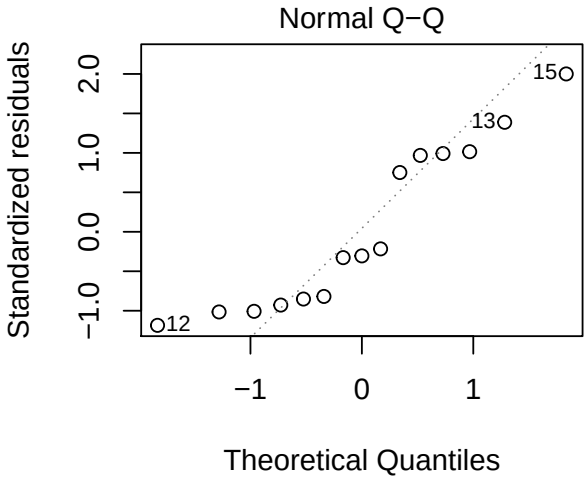
Q2: Interpret each of the two plots above.

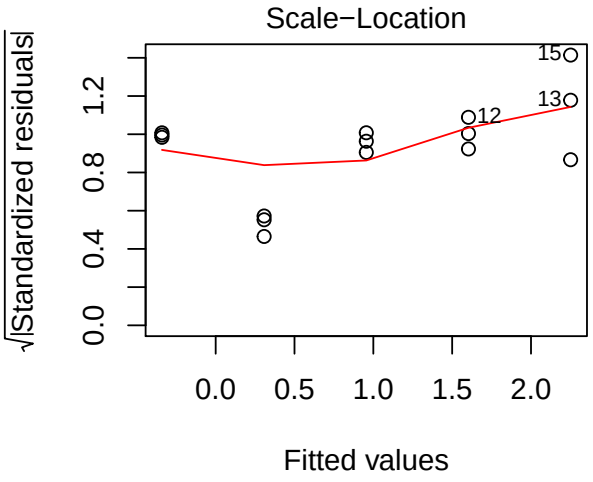
The first plot (Concentration VS Time) shows us an almost quadratic relationship between Concentration and Time. The datapoints are aligned on/depict the left side of the parabola (parabola is not shown in the picture, but one could picture it in mind). That said, the fitted line does not seem to address this relationship and missed a lot of datapoints resulting in a rather bad model. The relationship can also be considered logarithmic. Therefore, trying both, adding the quadratic term and the log transform (and picking the better out of those two models) would probably improve the model.

The second plot (Residuals VS Fitted Values) shows a non-constant spread. There are the signs of heteroscedasticity.

In both plots, we see a clearly non-linear relationship between the response and the predictor.

```
# The statement par (mfrow = c(1,2)) puts two plots side-by-side

par (mfrow = c(1,2))

# Plot the absolute residuals and the Normal Q-Q plot

plot (cofit, which=2:3)
```





Q3: Interpret the two plots above.

The first plot (Standardized Residuals VS Theoretical Quantiles) shows us a weak evidence that both sets of quantiles are coming from normal distributions. This is the case since the points are, to some degree, aligned

on the line, but again, this visual check is not a strong evidence.

The second plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows us a set of lines which are far from forming a straight line and/or being aligned on the same line. This plot shows that the residuals are not spread equally along the ranges of predictors and that the variance is not constant.

Fit a quadratic regression for Concentration vs Time.

```
# Insert R code to fit a quadratic regression model and summarize it.
hw4data$Timesq = (hw4data$Time - mean (hw4data$Time))^2

cofitQ <- lm(Concentration ~ Time + Timesq, data=hw4data)
summary(cofitQ)
```

```
##
## Call:
## lm(formula = Concentration ~ Time + Timesq, data = hw4data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2876 -0.1092  0.0261  0.1034  0.3572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.083905   0.109536  19.025 2.50e-10 ***
## Time        -0.324000   0.016433 -19.716 1.65e-10 ***
## Timesq       0.061429   0.006944   8.846 1.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.18 on 12 degrees of freedom
## Multiple R-squared:  0.9749, Adjusted R-squared:  0.9708
## F-statistic: 233.5 on 2 and 12 DF,  p-value: 2.473e-10
```

```
confint(cofitQ)
```

```
##                   2.5 %      97.5 %
## (Intercept)  1.84524527  2.32256426
## Time        -0.35980532 -0.28819468
## Timesq       0.04629806  0.07655908
```

Q4: Interpret the quadratic model results.

The mean response value when Time (in minutes) is 0 equals 2.084 ng/ml and is between 1.845 ng/ml and 2.323 ng/ml with 95% confidence.

The estimated mean response value (Time) is -0.324 ng/ml per minutes which changes between -0.360 ng/ml per minutes and -0.288 ng/ml per minutes, for any 1-unit increase in the predictor, with 95% confidence, holding all other predictors fixed.

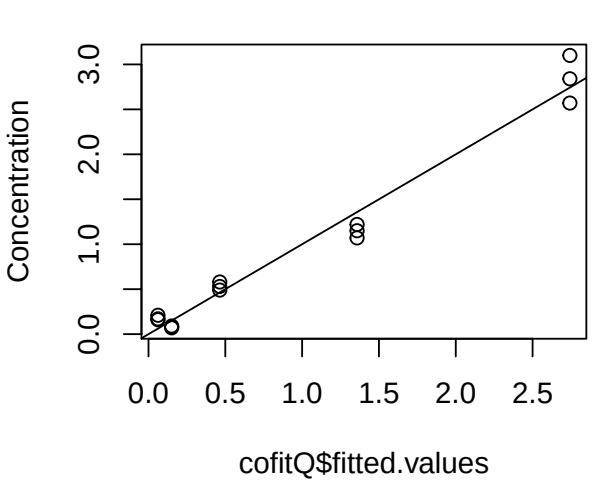
The estimated mean response value (Timesq) is 0.061 ng/ml per minutes which changes between 0.046 ng/ml per minutes and 0.077 ng/ml per minutes, for any 1-unit increase in the predictor, with 95% confidence, holding all other predictors fixed.

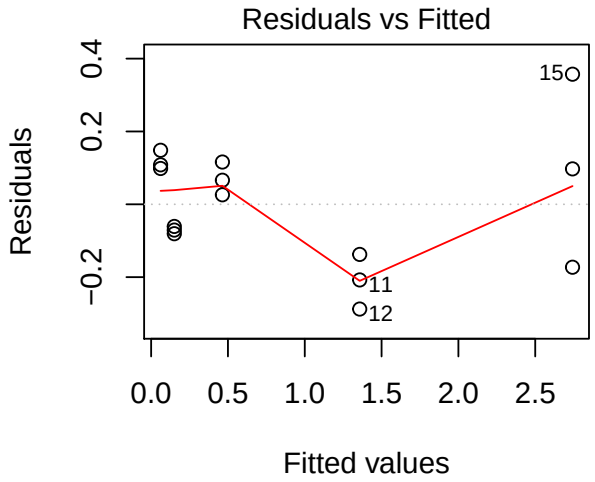
It should be noted that all three estimates Intercept, Time, Timesq are significant estimates with p-values of 2.50×10^{-10}, 1.65×10^{-10}, and 1.32×10^{-6} respectively.

The Adjusted R-squared value is 0.9708 which tells us that 97.08% of total variation in Concentration is explained by regressor variables. This is a lot better than 81.16% that we got in the previous model.

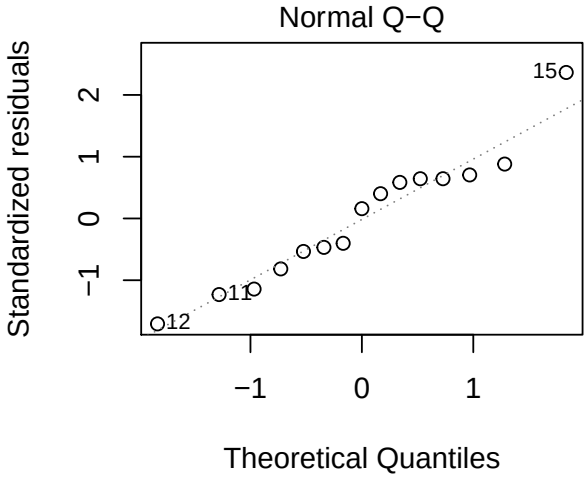
Residual standard error is 0.18 meaning that on average, predictions of the model are 0.18 ng/ml off from the actual value. This is a lot better than 0.4743 that we got in the previous model.

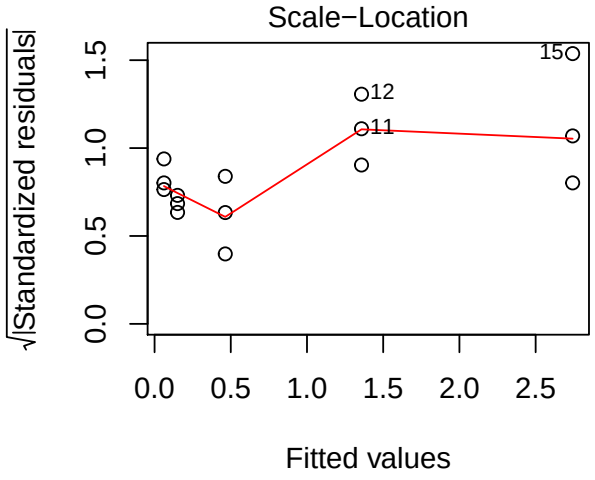
```
# Insert code to plot observed vs fitted Concentration values and
# make the first three residual plots by applying the plot() function
# to the fitted model object.  You should get these plots:  residuals vs
# fitted,  Normal Q-Q, and square root of absolute residuals vs fitted.
# This can be done in one or two code chunks.

par (mfrow = c(1,2))

# Plot the data and the fitted line
plot(Concentration ~ cofitQ$fitted.values, data=hw4data)
abline(0,1)

# Plot the residuals vs the fitted values
plot(cofitQ, which=1)
```






```
# Plot the absolute residuals and the Normal Q-Q
plot(cofitQ, which=2:3)
```





Q5: Interpret each of the plots for the quadratic fit.

The first plot (Concentration VS Time) shows us an almost linear relationship between Concentration and Time. That said, the fitted line does not seem to address this relationship well and missed a lot of datapoints

resulting in a rather bad model.

The second plot (Residuals VS Fitted Values) shows a rather non-constant spread, but it is a lot better than the initial model.

The third plot (Standardized Residuals VS Theoretical Quantiles) shows us a weak evidence that both sets of quantiles are coming from normal distributions. This is the case since the points are, to some degree, aligned on the line, but again, this visual check is not a strong evidence.

The fourth plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows us a set of lines which are very far from forming a straight line. This plot shows that the residuals are not spread equally along the ranges of predictors and that the variance is not constant. That said, this seems better than the plot and the relationship in first model.

```
# Fit a simple linear regression model using the transformed
# concentration data, Log10(Concentration) vs Time and summarize it.

hw4data$LogConc = log10 (hw4data$Concentration)
cofitL <- lm(LogConc ~ Time, data=hw4data)
summary(cofitL)
```

```
##
## Call:
## lm(formula = LogConc ~ Time, data = hw4data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.082958 -0.044421  0.006813  0.033512  0.085550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.654880   0.026181   25.01 2.22e-12 ***
## Time        -0.195400   0.004557  -42.88 2.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04992 on 13 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9924
## F-statistic:  1838 on 1 and 13 DF,  p-value: 2.188e-15
```

```
confint(cofitL)
```

```
##                  2.5 %     97.5 %
## (Intercept)  0.5983196  0.7114399
## Time        -0.2052462 -0.1855544
```

Q6: Interpret the regression model results above.

The mean response value when Time (in minutes) is 0 equals 0.655 log(ng)/log(ml) and is between 0.598 log(ng)/log(ml) and 0.711 log(ng)/log(ml) with 95% confidence.

The estimated mean response value is -0.195 log(ng)/log(ml) per minutes which changes between -0.195 log(ng)/log(ml) per minutes and 0.005 log(ng)/log(ml) per minutes, for any 1-unit increase in the predictor, with 95% confidence.

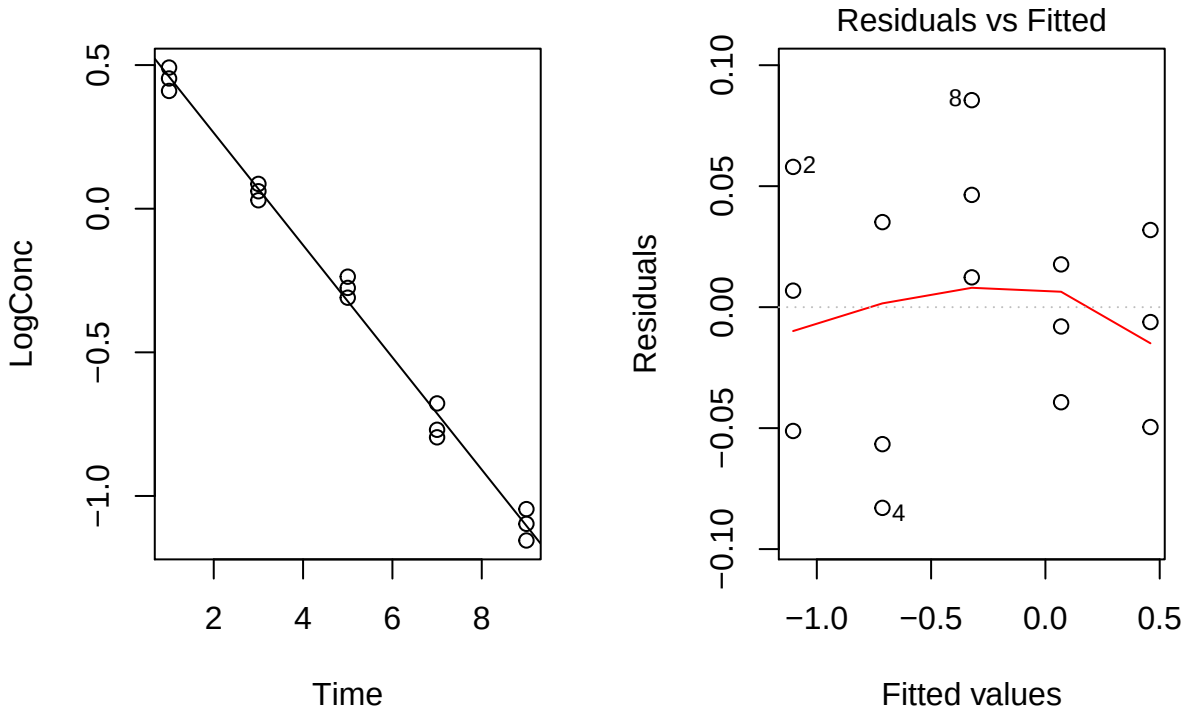
It should be noted that both intercept and slope are significant estimates with p-values of 2.22×10^{-12} and 2.19×10^{-15} respectively.

The Multiple R-squared value is 0.933 which tells us that 93.3% of total variation in Log Concentration is explained by Time. This is a bit lower than the Adjusted R-squared in the model where we had a quadratic

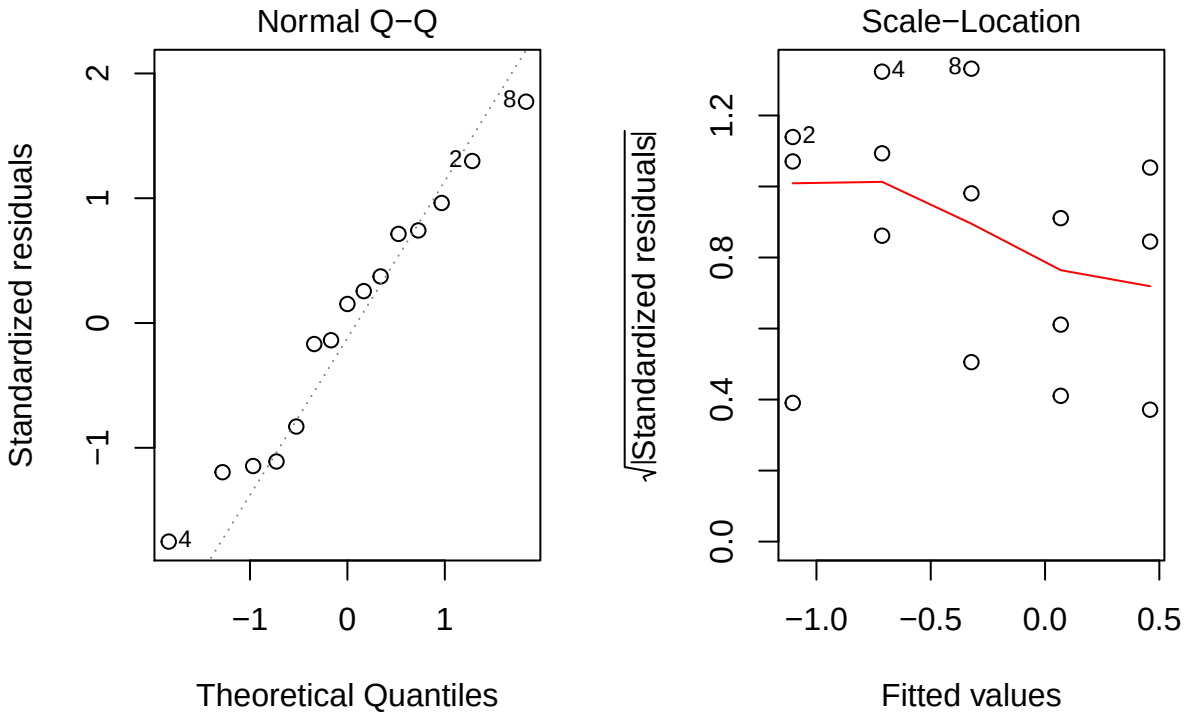
term (96.08%), but that said, Adjusted R-squared and Multiple R-squared are different and the models are fundamentally different (we have multiple linear regression VS simple linear regression).

Residual standard error is 0.050 meaning that on average, predictions of the model are 0.050 ng/ml off from the actual value. This is a big (almost tenfold) improvement over the quadratic term model value of 0.4743 ng/ml.

```
par (mfrow = c(1,2))

# Plot Log Concentration vs Time and add the fitted line for that model
plot(LogConc ~ Time, data=hw4data)
abline(cofitL)

# Make the first 3 residual plots
plot(cofitL, which=1)
```




```
plot(cofitL, which=2:3)
```



Q7: Interpret the plots above.

The first plot (Concentration VS Time) shows a strong linear relationship between LogConc and Time. The fit is also very good!

The second plot (Residuals VS Fitted Values) shows a constant spread and no signs of heteroscedasticity.

The third plot (Standardized Residuals VS Theoretical Quantiles) shows us a fairly strong evidence that both sets of quantiles are coming from normal distributions. This is the case since the points are, to some degree, aligned on the line.

The fourth plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows us a set of lines which are rather close to forming a straight line suggesting that the residuals are spread equally along the ranges of predictors and that the variance is practically constant.

Overall, the logarithmic model seems to be the best, followed by the quadratic model, and finally, the initial model.