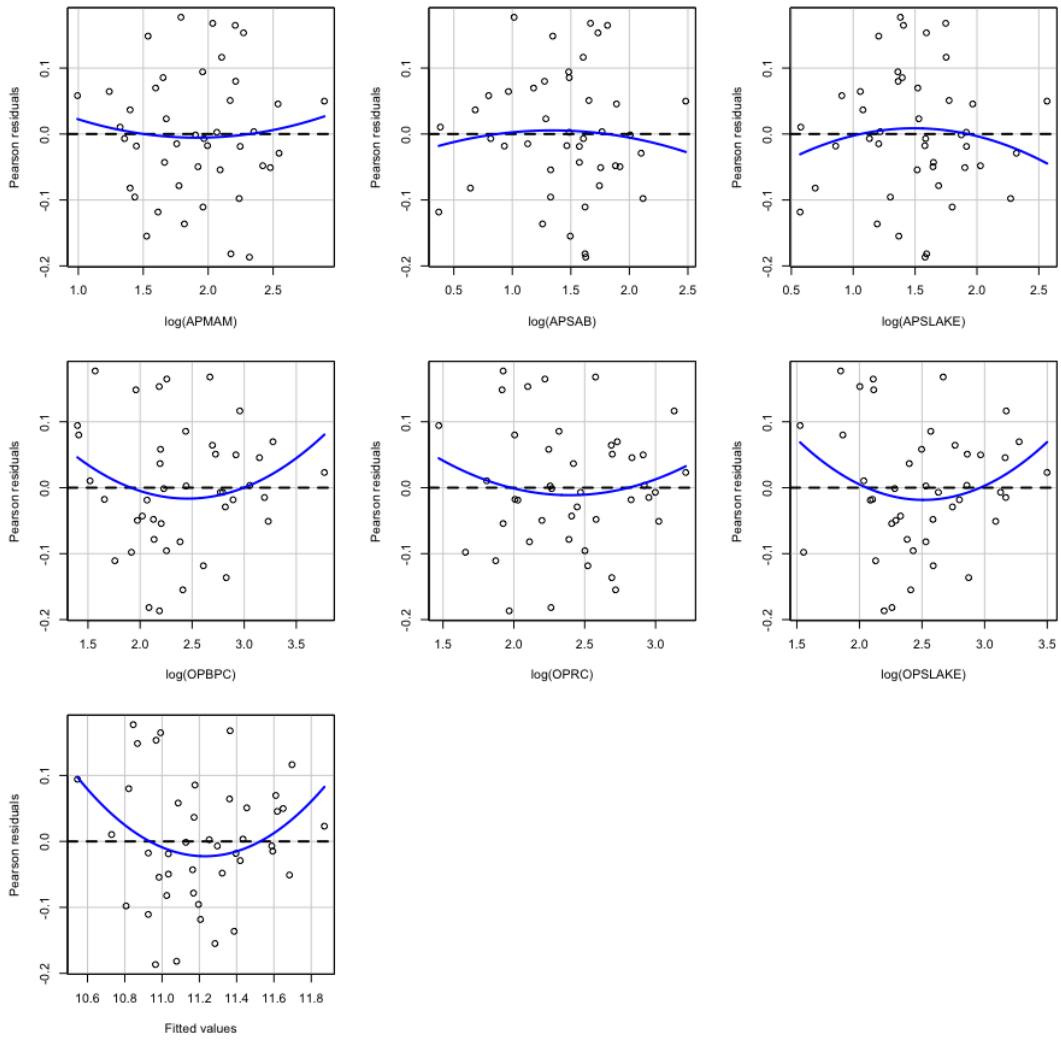
Exercise 1: Do problem 9.8. Draw residual plots for the mean function described in Problem 8.3.4 for the California water data, and comment on your results. Test for curvature as a function of fitted values.

Solution:
Recall that problem 8.3.2 described the following model,

$$\log(BSAAM) \sim \log(APMAM)+\log(APSAB)+\log(APSLAKE)+\log(OPBPC)+\log(OPRC)+\log(OPSLAKE)$$

Creating the model in r, and calling the residualplots() we get the following figure. We can see that the residual plots for each first order predictor do not exhibit significant curvature. Each plot looks fairly random, and when we consider the tests on the second order predictors that are produced by residualplots(), at an $\alpha = .05$ significance level we get that our residual plots show little curvature.

Figure 1: Residual Plots for 8.3.2 model.



Code:

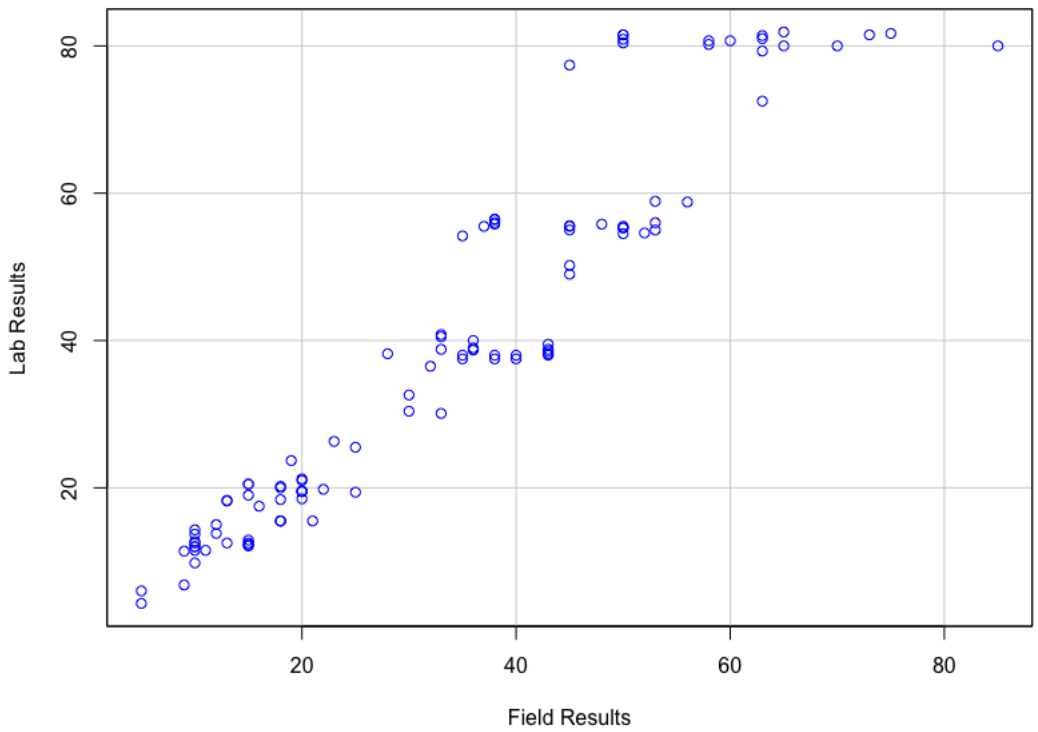
```
> df <- water

> WaterModel <- lm(log(BSAAM) ~ log(APMAM)+
                                log(APSAB)+
                                log(APSLAKE)+
                                log(OPBPC)+
                                log(OPRC)+
                                log(OPSLAKE), data = df)

> residualPlots(WaterModel)

              Test stat Pr(>|Test stat|)
log(APMAM)       0.4499           0.65553
log(APSAB)      -0.4647           0.64502
log(APSLAKE)    -0.8525           0.39976
log(OPBPC)       1.3848           0.17487
log(OPRC)        0.8387           0.40735
log(OPSLAKE)     1.6295           0.11217
Tukey test       1.8386           0.06597 .
```

Exercise 2: Do problem 9.3, and skip part 9.3.3. This example compares in=field ultrasonic measurements of depths of defects, Field, in the Alaska oil pipeline with measurements of the same defects in a laboratory, Lab. the lab measurements were done in six different batches, in the variable Batch. The goals is to decide if the field measurement san be used to predict the more accurate lab measurements. The lab measurement is the response variable and the field measurement is the predictor variable. The data are from National institute of Science and Technology.

9.3.1 Draw the scatterplot of Lab vs Field, and comment on the applicability of a simple linear regression model.

Solution:
Using r we get the following scatterplot(code is in next part). Here a simple linear regression would likely do a bad job at explaining the relationship between Lab and Field data. In an early module we discussed how when a scatterplot has a fan shape that is an indication of non-constant variance of an SLR model. We can see that as our lab results grow in size so does the spread of the data resulting in the fan shape that we discussed previously.

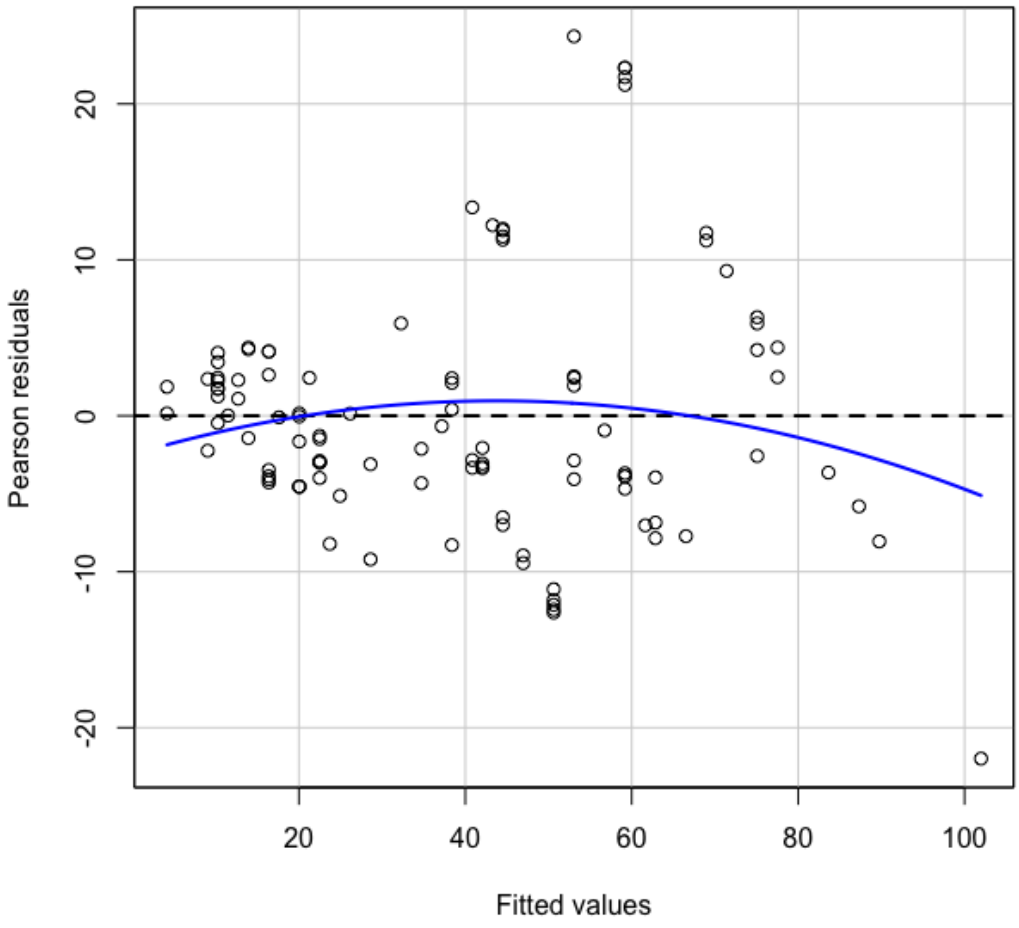
Figure 2: Lab vs Field Scatterplot



9.3.2 Fit the simple regression model, get the residual plot, and summarize. Explain why the plot suggests non-constant variance and provide a test for non-constant variance.

Solution:
Fitting the SLR model and producing the residual plot in r we get the following. The residual plot shows a left to right fanning pattern (like $' <'$) which suggests that we are dealing with a non-constant variance. With a p-value of 5.3499e-08 the Bruesch-Pagan test confirms our idea that we are dealing with non-constant variance.

Figure 3: Residual Plots for Pipeline SLR



Code:

```
df <-pipeline
scatterplot(df$Field, df$Lab, regLine = FALSE,
            boxplots = FALSE,
            smooth = FALSE,
            xlab = 'Field Results',
            ylab = 'Lab Results')


SLR_Pipline <- lm(Lab ~ Field, data = df)
residualPlots(SLR_Pipline, terms = ~-1)
           Test stat Pr(>|Test stat|)
Tukey test   -1.3025           0.1927

ncvTest(SLR_Pipline)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 29.58568, Df = 1, p = 5.3499e-08
```

Exercise 3: Using the data and model fit in problem 9.8, do the following,

1. Perform a test for outliers and interpret the test's results.

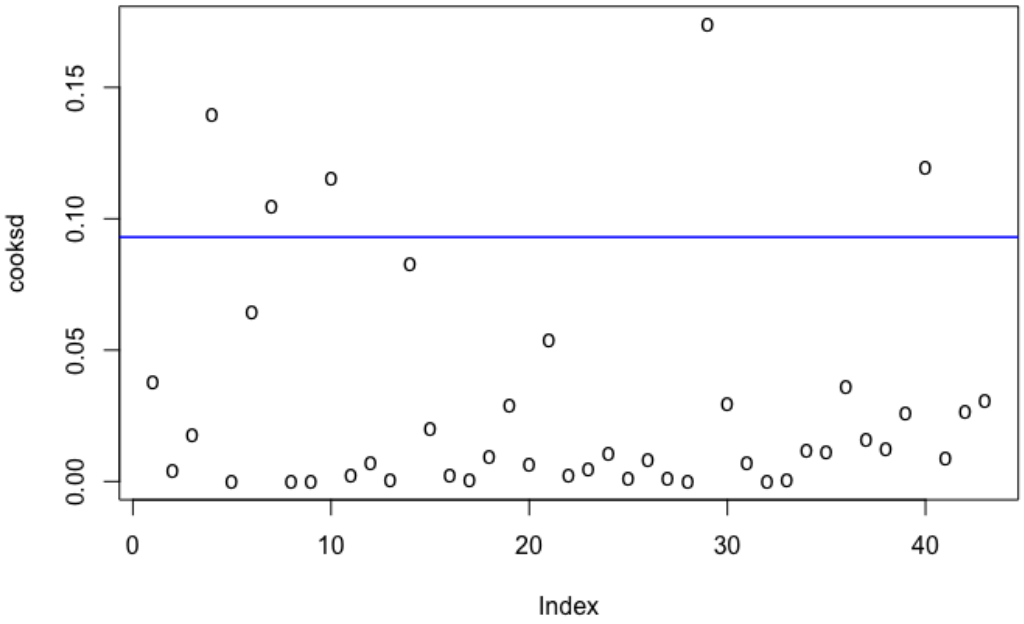
Solution:
Using the outlierTest() function and computing the t-statistic for all studentized residuals in r we get that there are no outliers even if we adjust our significance by using the Bonferroni p-value,
Code:

```
df <- water
WaterModel <- lm(log(BSAAM) ~ log(APMAM)+
                      log(APSAB)+
                      log(APSLAKE)+
                      log(OPBPC)+
                      log(OPRC)+
                      log(OPSLAKE), data = df)

residuals <- rstudent(WaterModel)

# Computing 2 sided p-value had 43 data, and 6 paramaters
> 2*pt(t(residuals), 43 - 6 - 2)
       1         2         3          4         5         6         7
1.626718 0.6573949 0.3260013 0.08011789 0.9422671 1.896411 0.17365
       8         9        10         11        12        13        14
0.9885036 1.033643 1.930214 1.386043 0.5744184 1.092296 0.047476
       15        16        17         18        19        20        21
0.1681711 0.7600528 0.8572238 0.6062299 0.3746106 1.367178 1.8814
      22        23        24         25        26        27        28
1.198641 1.302774 0.4246441 0.8457937 1.490373 0.8492326 1.0216
       29        30        31         32        33        34        35
0.2074365 1.683469 0.5947018 0.9420563 0.8810112 1.463623 1.4112
      36       37        38        39        40        41        42
1.778421 1.90728 1.617154 1.557578 1.947318 0.5983426 0.065578
       43
0.2472316

# Testing with outlierTest()
outlierTest(WaterModel)
    No Studentized residuals with Bonferroni p < 0.05
    Largest |rstudent|:
        rstudent unadjusted p-value Bonferroni p
    14 -2.054273           0.047477           NA
```

2. Determine whether any observations are highly influential, as measured by Cook's D.

Solution:
Using the cooks.distance() function in r we can find which observations have high influence. Since there are conflicting practices for determining the significance of the Cook's D benchmark we will consider both the greater than 1 and greater than 4/n criteria. From r, we got that with the 4/n criteria there are 5 highly influential observations that seem to be spread out evenly throughout the data. With the greater than 1 criteria we get that there are no influential observations.

Figure 4: Plot of Cook's D Values, With 4/n Criteria



Code:

```
cooksd <- cooks.distance(WaterModel)
plot(cooksd, pch="o")
abline(h = 4/length(cooksd), col="blue")
# Index of influential observations with 4/n criteria
Influential_index <- as.numeric(names(cooksd)[(cooksd > (4/length(cooksd)))])
    [1]  4  7 10 29 40
```

3. Perform a Durbin-Watson test of independence in the residuals and interpret the test's result.

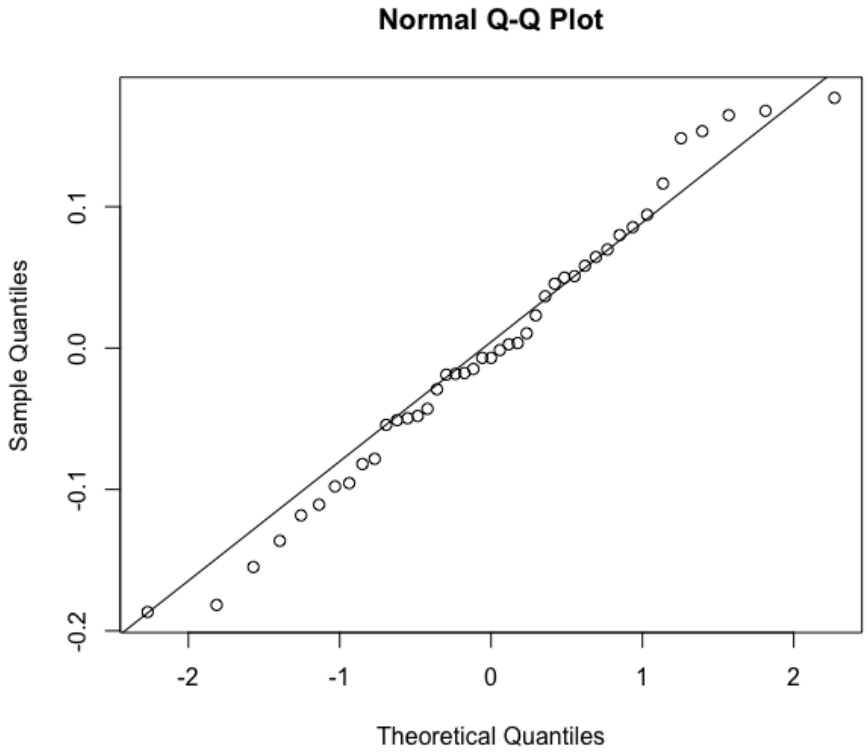
Solution:
Performing the Durbin-Watson test in r, we get a p-value of 0.034. On the $\alpha = .05$ we reject the null hypothesis and concluded that there does exists some autocorrelation among the residuals of the model. With a p-value so close to our significance threshold we might be able to get away with using the model, however this results warrants further analysis. We could try to refit the model with different(less) predictors or we could try some transformations on the data.
Code:

```
dwt(WaterModel)
 lag Autocorrelation D-W Statistic p-value
   1       0.2843484      1.381223   0.034
 Alternative hypothesis: rho != 0
```

4. Obtain a normal probability plot and perform a a Shapiro Wilk test of normality in the residuals. Interpret the test's results and the plot.

Solution:
Producing the normal probability plot and performing the Shapiro-Wilk test we get that the residuals are normally distributed. The normal probability plot shows some group of residuals below the fitted line in the beginning and above the fitted line at the end. This suggests that the probability distribution for our residuals had larger tails than that of a normal or t-distributions. With a p-value of 0.6659, we reject the null and the Shapiro-Wilk test tells us that this difference is insignificant on the $\alpha = .05$ level.

Figure 5: Normal Probability Plot for WaterModel Residuals



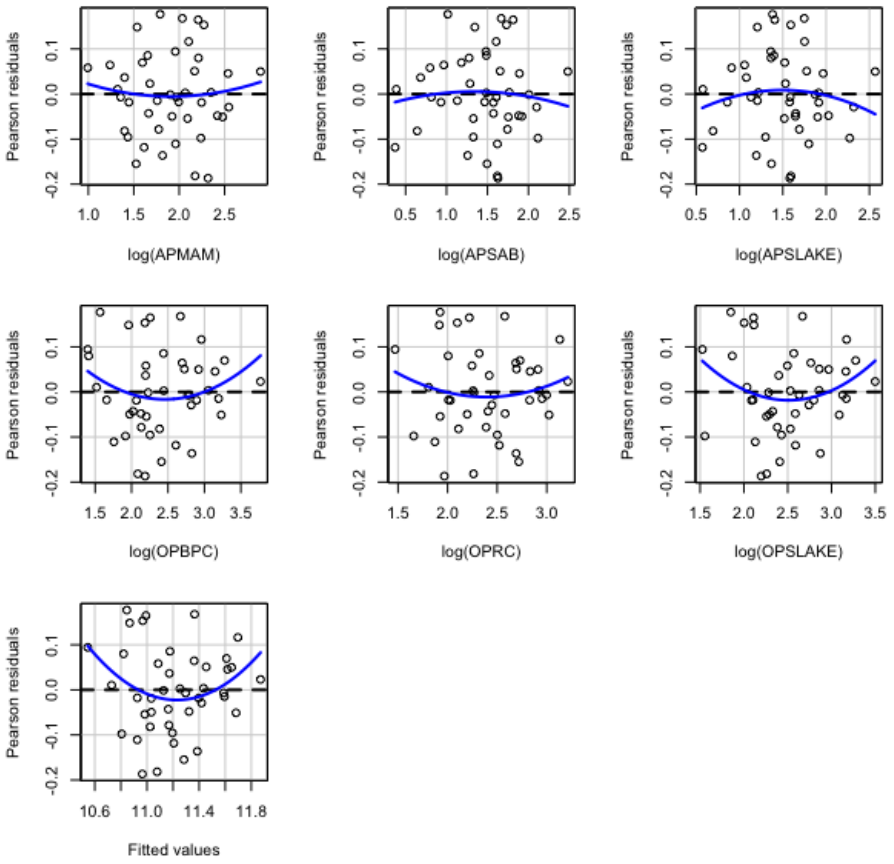
Code:

```
shapiro.test(residuals(WaterModel))

        Shapiro-Wilk normality test

data:  residuals(WaterModel)
W = 0.98047, p-value = 0.6659
```

5. Perform a test for non-constant variance in the residuals vs. fitted values plot. Interpret the test's results.

Solution:
Using the residualPlots() command we can obtain the residual plots for each predictor, as well as the fitted values of the model. We also get the p-values for the significance of second order predictors, as well as Tukey's test for non-constant variance. Tukey's test shows us that with a p-value of 0.06597, on the $\alpha = .05$ level the residuals demonstrate constant variance.

Figure 6: Residual Plots for WaterMode



Code:

```
> residualPlots(WaterModel)
              Test stat  Pr(>|Test stat|)
log(APMAM)       0.4499           0.65553
log(APSAB)      -0.4647           0.64502
log(APSLAKE)    -0.8525           0.39976
```

```
log(OPBPC)      1.3848          0.17487
log(OPRC)       0.8387          0.40735
log(OPSLAKE)    1.6295          0.11217
Tukey test      1.8386          0.06597 .
```

6. Summarize all your finding from the diagnostics of this model and make recommendations about whether this model's results should be trusted.

Solution:
Of the tests that we performed we found there to be several influence points using the 4/n criteria for the Cook's D test. We also found, from the Durbin-Watson test that there does exists some autocorrelation in the residuals of the model. In that problem we suggested that we could try transforming the data or fitting the data with different predictors. In any case, I would encourage further analysis before trusting this model, maybe through PCA we can test out different predictors and transformations to get a better model.