HW3 Solutions

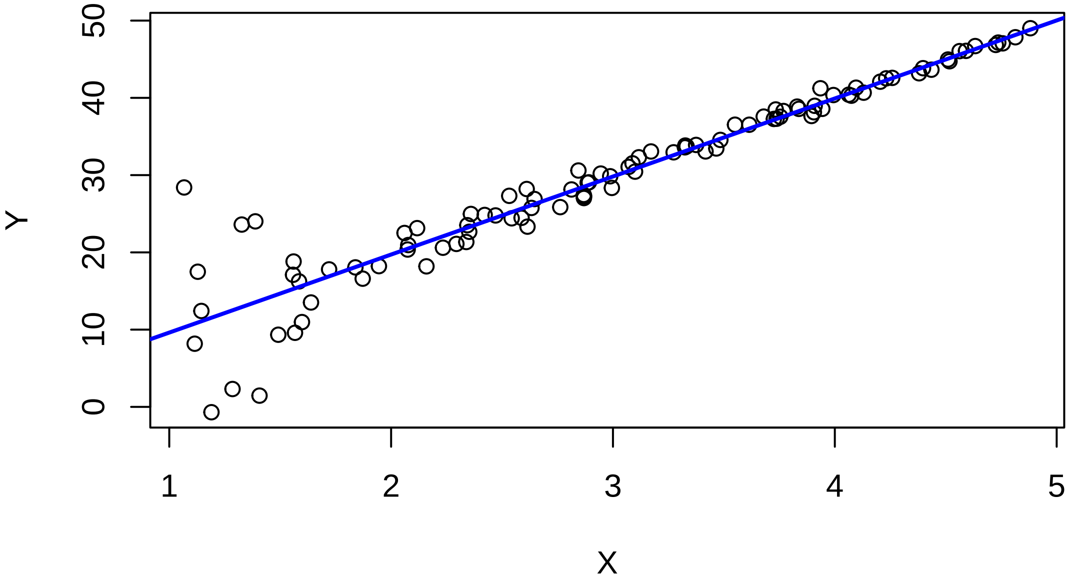
Based on the work of Robert P.

Problem 3.2

It is perfectly fine if you draw the figure. Here we shous how to do it using simulated data. Let's simulate some data so that we obtain the desired residual plots:

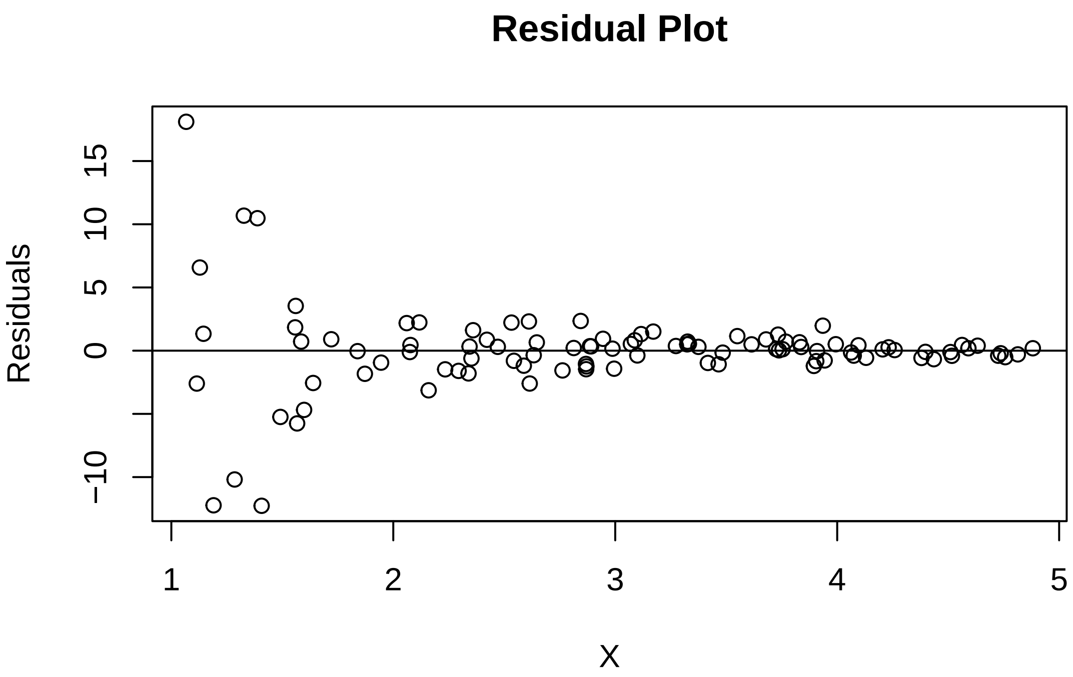
Part (1)

```
set.seed(2894219)

# Generate x from Uniform distribution:
x_vals <- runif(100, 1, 5)
# Generate y from Normal dist. such that the spread decreases as x increases:
y_vals <- rnorm(100, mean = 10*x_vals, sd = 10/x_vals^2)

lm_sim1 <- lm(y_vals ~ x_vals)

plot(x_vals, y_vals, xlab = "X", ylab = "Y")
abline(lm_sim1, col = "blue", lwd = 2)
```



So here is the residual plot which shows the error variance decreasing as X increases:

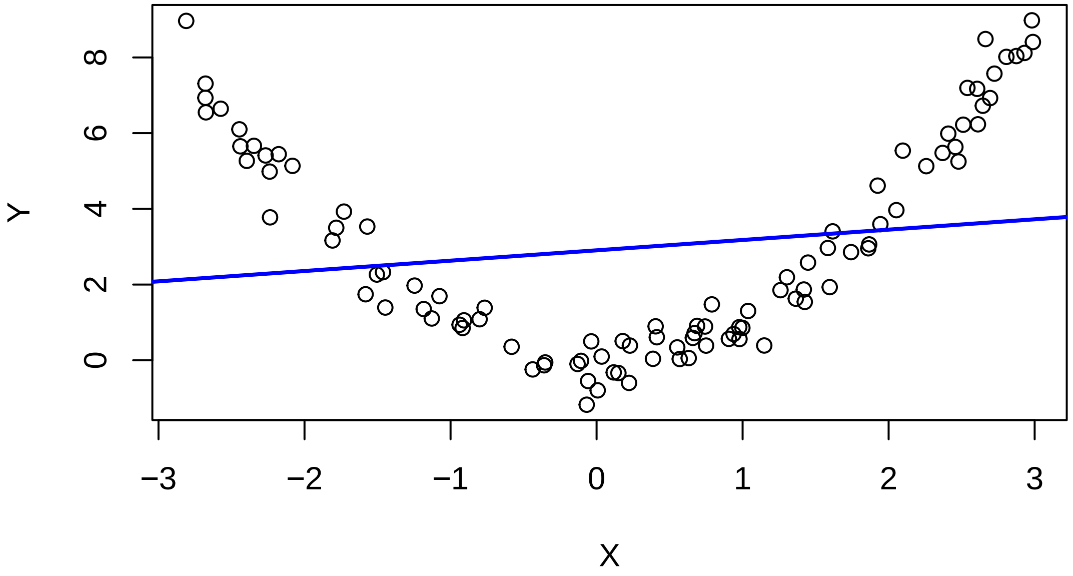
```
plot(x_vals, lm_sim1$residuals, xlab = "X", ylab = "Residuals",
     main = "Residual Plot")
abline(h=0)
```



Part (2)

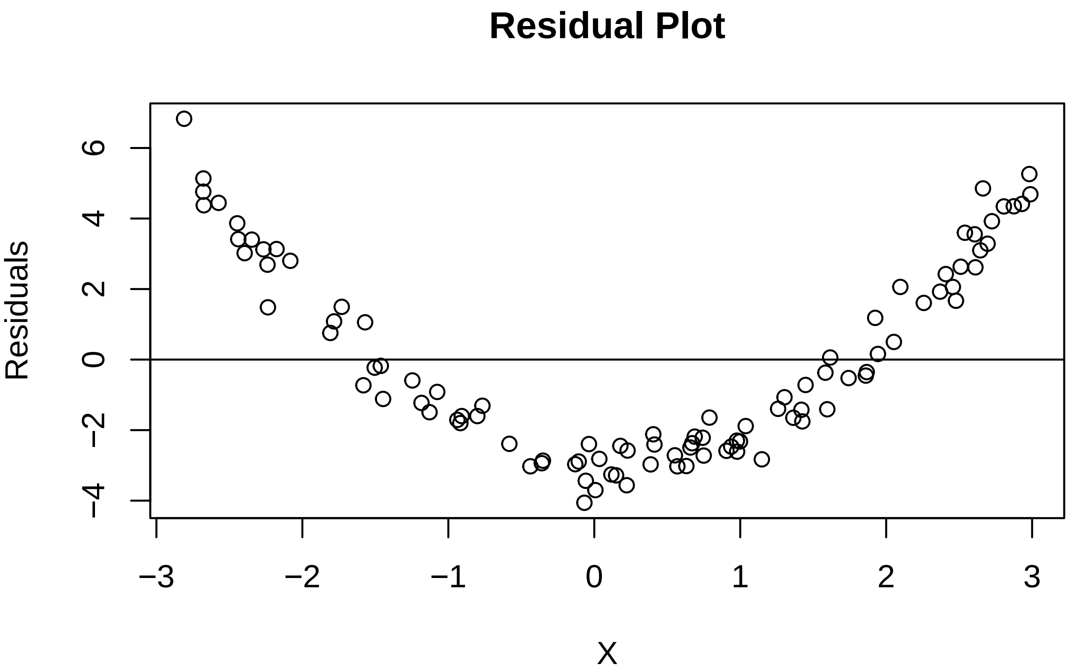
Now, let's simulate data so that the true regression function is U-shaped:

```
set.seed(2752902)

x_vals <- runif(100, -3, 3)
# Square the x-values and add some random noise:
y_vals <- x_vals^2 + rnorm(100, 0, 0.5)

lm_sim2 <- lm(y_vals ~ x_vals)

plot(x_vals, y_vals, xlab = "X", ylab = "Y")
abline(lm_sim2, col = "blue", lwd = 2)
```



Here is the residual plot for our linear regression model:

```
plot(x_vals, lm_sim2$residuals, xlab = "X", ylab = "Residuals",
     main = "Residual Plot")
abline(h=0)
```

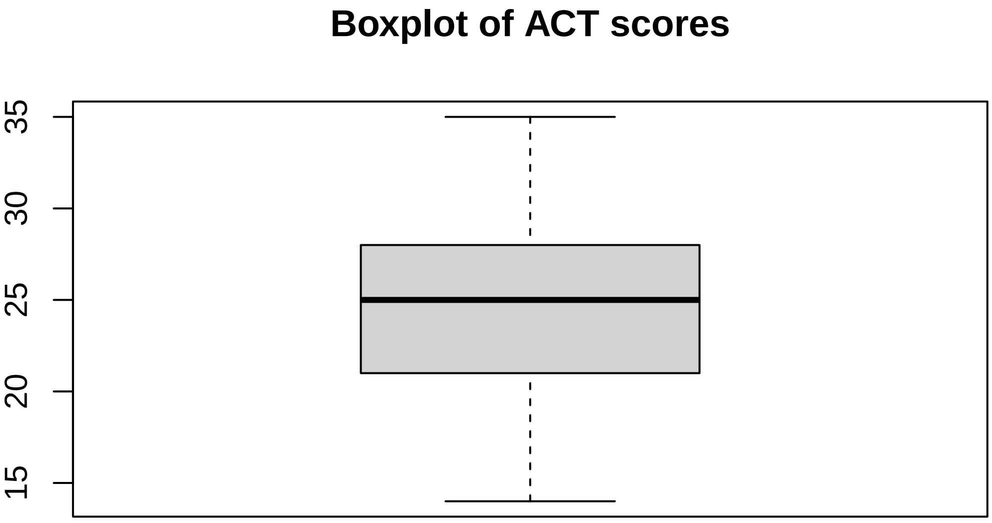


Problem 3.3

```
gpa.dat <- read.table("./BookDataSets/Chapter  3 Data Sets/CH03PR03.txt", header = F)
gpa <- gpa.dat[,1]
act <- gpa.dat[,2]
```

Part (a)

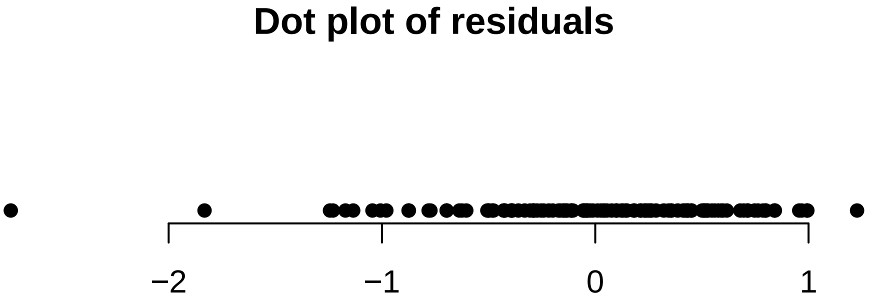
```
boxplot(act, main = "Boxplot of ACT scores")
```



This seems like a very standard box plot. It appears that ACT scores are roughly symmetric and do not appear to have any outliers.

Part (b)

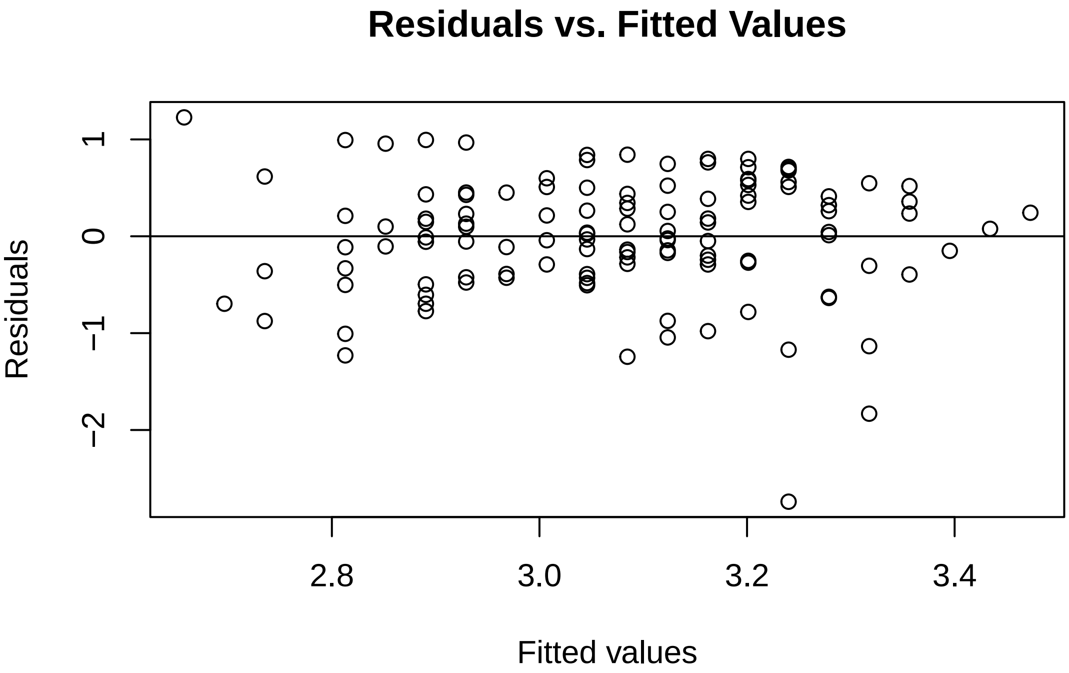
```
lm.gpa <- lm(gpa ~ act)
gpa.res <- lm.gpa$residuals

stripchart(gpa.res, main = "Dot plot of residuals", frame.plot = F,
           pch = 16, axes = F)
axis(1, pos = 0.95)
```



We can use this dot plot of residuals to detect outliers. According to the plot, the vast majority of residual values are between -1 and 1, but there are a few residuals with magnitudes that are bigger than 1. There is one particularly large negative residual.

Part (c)

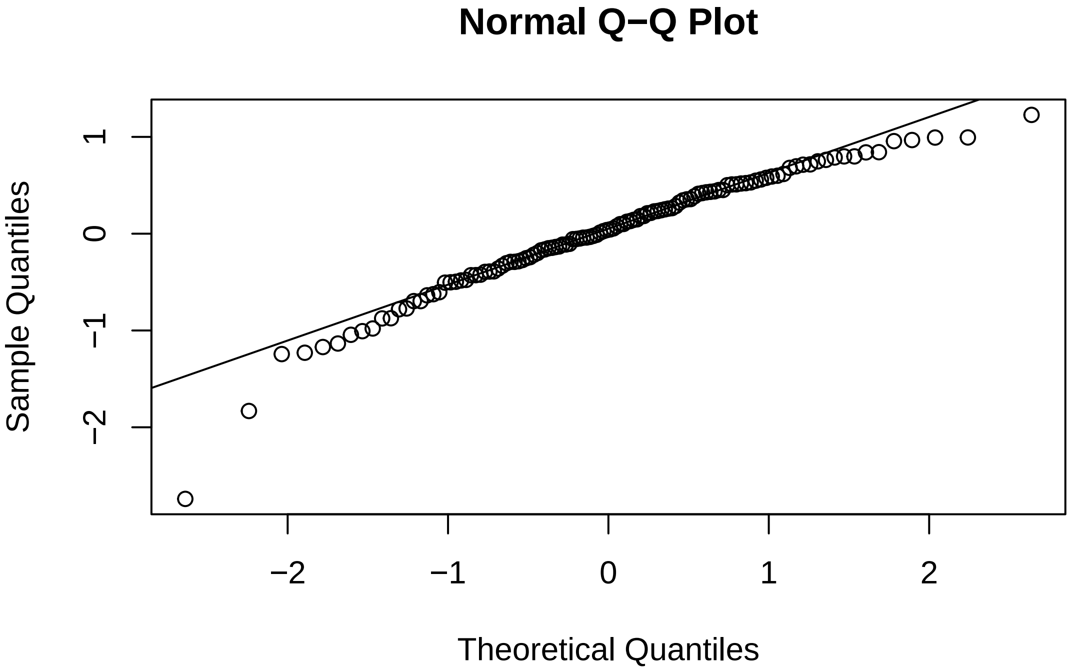
```
gpa.fitted <- lm.gpa$fitted.values

plot(gpa.fitted, gpa.res, xlab = "Fitted values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values")
abline(h=0)
```



Overall, there appears to be somewhat of an upside-down U shape to the residuals, which might indicate a violation of the linearity assumption. There are also a few rather large residual values, which may indicate that there are outliers. In addition, the constant variance assumption may be slightly violated, but this may just be due to the presence of outliers. However, the degree of the violation (if any) is not large, so we should be safe in fitting a linear regression model.

Part (d)

```
qqnorm(gpa.res)
qqline(gpa.res)
```



You also use the fucntion of plot(gpa.fitted) which will generate four graphs.

We see that the majority of points fall on the line, which may indicate that the normality assumption is probably not severely violated. However, at the left and right extremes, the points do appear to deviate from the line. This is likely due to the outliers in our data.

```
res.expected <- qqnorm(y = gpa.res)$x
```

```
cor(gpa.res, res.expected)
```

```
## [1] 0.9744497
```

You can use the R example for Chapter 2 GPA problem to get the expected residuals or you can extract the information from qqnorm command.

The correlation coefficient between the ordered residuals and their expected values under normality is 0.974. Looking at Table B.6 with $\alpha = 0.05$ and $n = 100$, we see that the critical value is 0.987. Since 0.974 is less than the critical value, we cannot conclude that the error terms are reasonably normally distributed. As mentioned earlier, this is probably due to the presence of outliers in our data.

Part (e)

```
gpa.dat$res <- gpa.res
gpa.low <- gpa.dat[gpa.dat$V2 <26,]
gpa.high <- gpa.dat[gpa.dat$V2 >=26,]

d_low <- abs(gpa.low$res - median(gpa.low$res))
d_high <- abs(gpa.high$res - median(gpa.high$res))

# two-sample t-test
t.test(d_low, d_high, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  d_low and d_high
## t = -0.89674, df = 118, p-value = 0.3717
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.21994668  0.08283519
## sample estimates:
## mean of x mean of y
## 0.4379603 0.5065161
```

```
# Critical value
crit_val <- qt(p = 1-0.01/2,df = 118)
crit_val
```

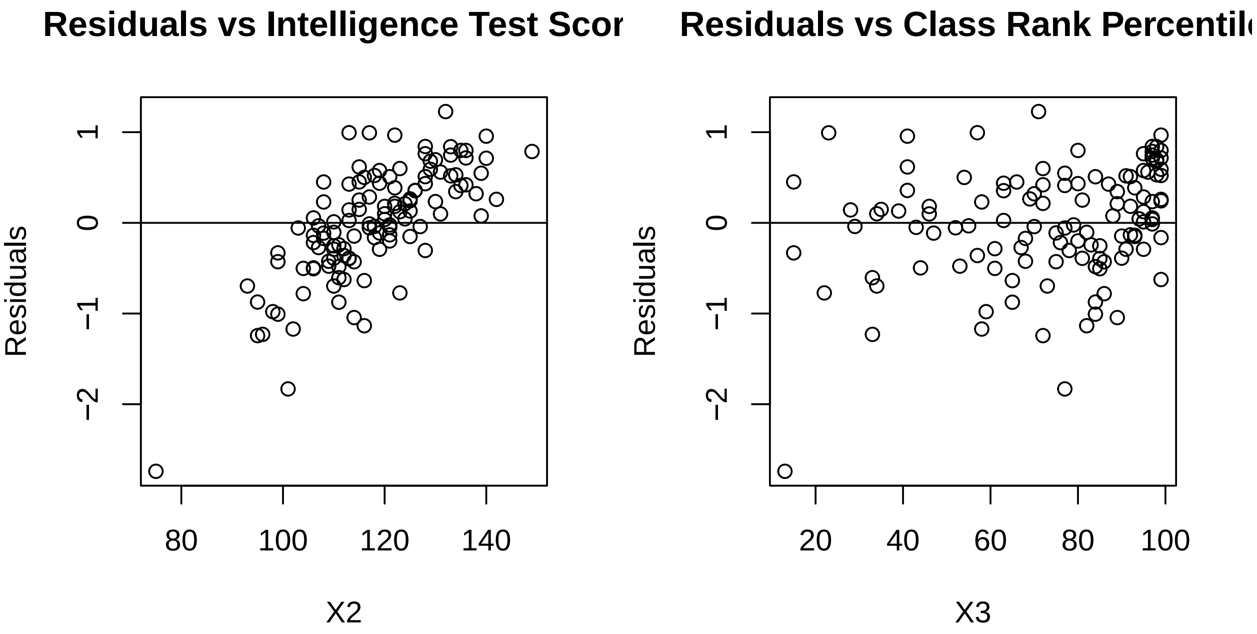
```
## [1] 2.618137
```

The decision rule of the Brown-Forsythe test is: If the absolute value of the t-statistic is larger than the critical value (2.62), we reject the null hypothesis. Otherwise, we fail to reject.

We see that the magnitude of our t-statistic is 0.897, which is smaller than the critical value. Thus, we fail to reject the null hypothesis and conclude that we do not have significant evidence that the error variance is not constant. Thus, it appears that the constant variance assumption is not violated.

Part (f)

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

plot(gpa.dat$V3, gpa.res, main="Residuals vs Intelligence Test Score", xlab="X2", ylab="Residuals")
abline(h=0)

plot(gpa.dat$V4, gpa.res, main="Residuals vs Class Rank Percentile", xlab="X3", ylab="Residuals")
abline(h=0)
```



We should include Intelligence test score in our model since the residual plot indicates a linear relationship between the residuals of our current model and X_2. The residuals and X_3 are only slightly correlated, and the residuals does not appear to vary systematically with the level of X_3. Thus, it appears that adding X_3 will be less useful than adding X_2.

Problem 3.15

```
solution.dat <- read.table("./BookDataSets/Chapter  3 Data Sets/CH03PR15.txt", header = F)
names(solution.dat) <- c("concentration", "time")
```

Part (a)

```
lm.solution <- lm(concentration ~ time, data = solution.dat)

summary(lm.solution)
```

```
##
## Call:
## lm(formula = concentration ~ time, data = solution.dat)
##
## 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
```

Part (b)

The null and alternative hypotheses are:

$$H_0 : E[Y|X] = \beta_0 + \beta_1 X$$
$$H_1 : E[Y|X] \neq \beta_0 + \beta_1 X$$

The decision rule is that we reject the null if $F^* > F_{0.975}(3, 10) = 4.83$. Otherwise we fail to reject.

```
qf(0.975, 3, 10)
```

```
## [1] 4.825621
```

```
lm.full <- lm(concentration ~ as.factor(time), data = solution.dat)

anova(lm.solution, lm.full)
```

```
## Analysis of Variance Table
##
## Model 1: concentration ~ time
## Model 2: concentration ~ as.factor(time)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1     13 2.9247                                  
## 2     10 0.1574  3    2.7673 58.603 1.194e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see from the R output above that $F^* = 58.60$, which is greater than the critical value. Thus, we reject the null hypothesis and conclude that the regression function is not linear, and so the linear regression model is not a good fit.

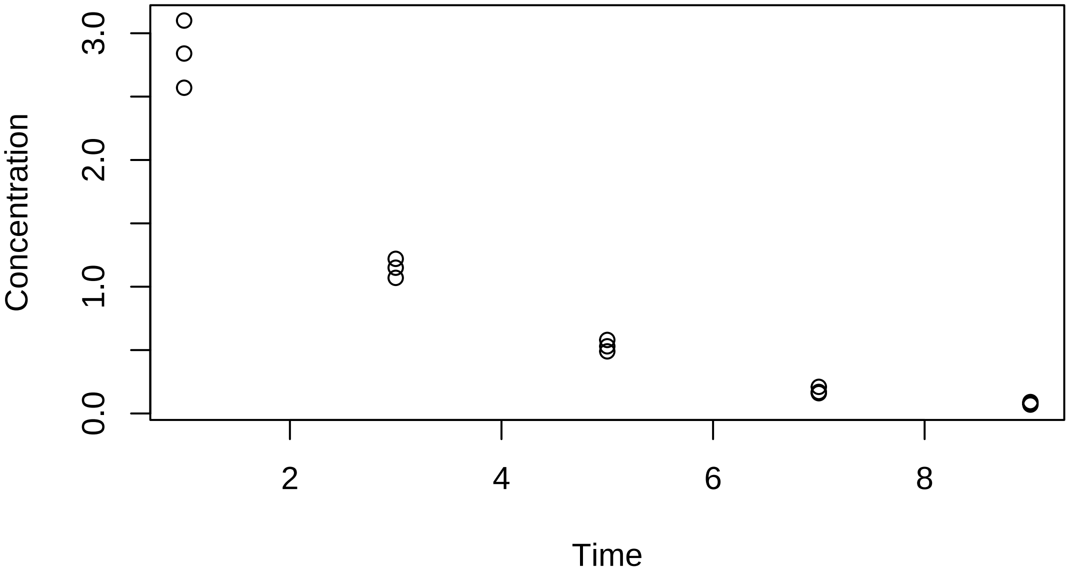
Part (c)

No, if we reject the null hypothesis in the test in Part (b), we can only conclude that the linear regression model is not a good fit. However, it gives us no information on what regression function is actually appropriate. The alternative includes all regression functions other than a linear one. We need to do more work (such as examining the residuals) to see what regression functions may be appropriate.

Problem 3.16

Part (a)

```
plot(solution.dat$time, solution.dat$concentration, xlab = "Time",
     ylab = "Concentration")
```



A log transformation of the Y variable seems like it might achieve constant variance and linearity.

Part (b)

```
library(MASS)

# Get the log-likelihood for the chosen lambda values
with(solution.dat, boxcox(concentration~time, lambda=seq(-0.2,0.2,0.1),
                          plotit=FALSE))
```

```
## $x
## [1] -0.2 -0.1  0.0  0.1  0.2
##
## $y
## [1]  4.558814  9.368724 13.210948 12.307766  7.694723
```

Of the given values, the likelihood is maximized when $\lambda = 0$. This corresponds a log transformation on Y, which is what we concluded in Part (a).

Part (c)

```
lm.solution.log <- lm(log10(concentration)~time, data = solution.dat)
lm.solution.log
```

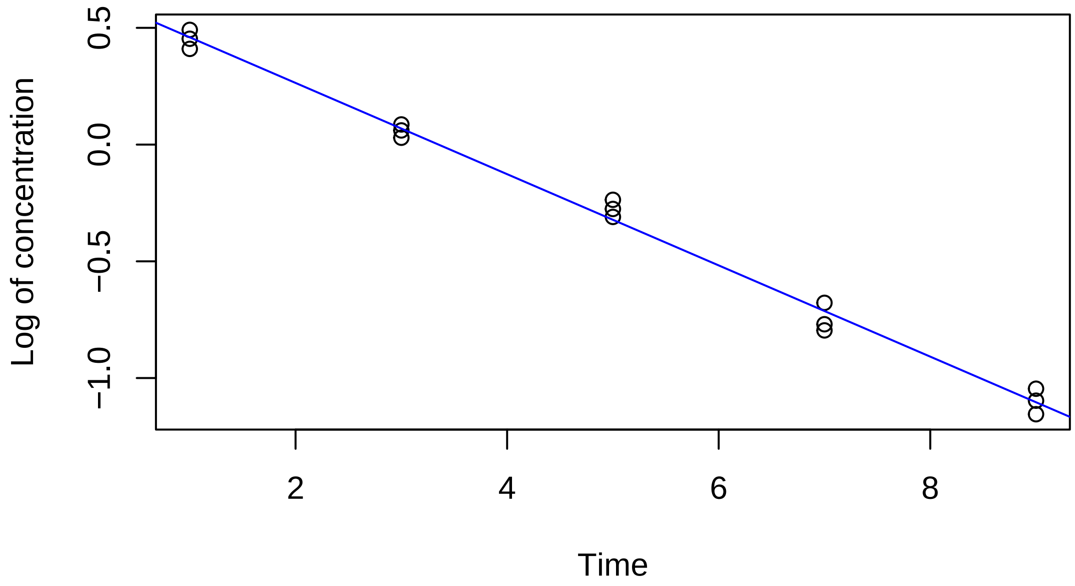
```
##
## Call:
## lm(formula = log10(concentration) ~ time, data = solution.dat)
##
## Coefficients:
```

```
## (Intercept)        time
##      0.6549     -0.1954
```

The estimated regression function is $E(Y') = 0.655 - 0.195 \cdot \text{time}$.

Part (d)

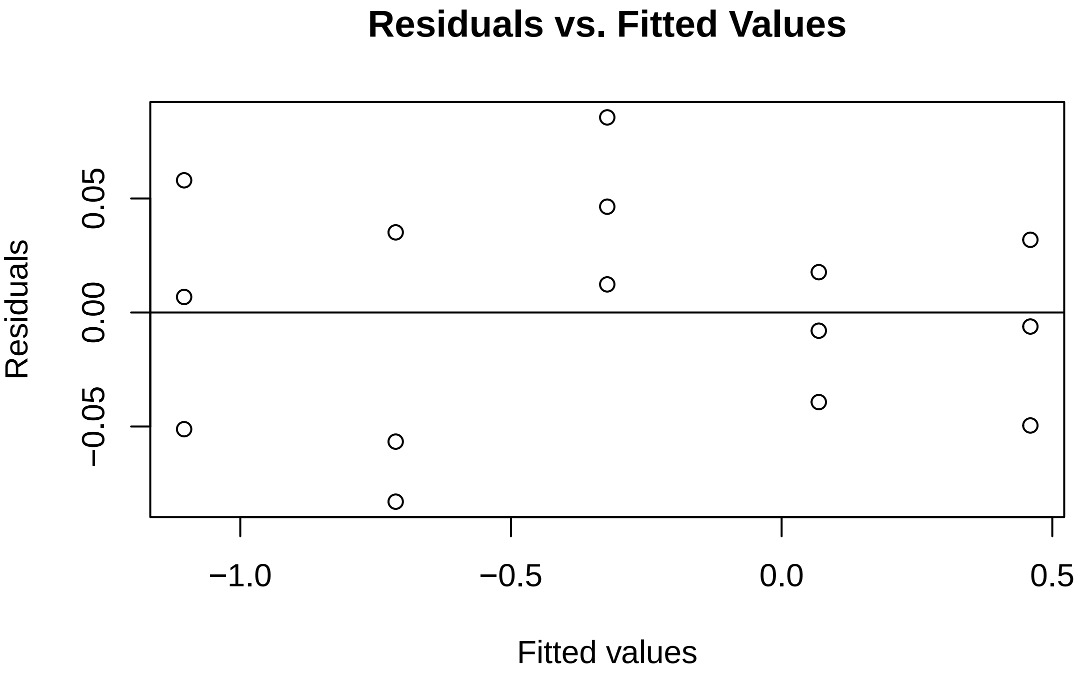
```
plot(solution.dat$time, log10(solution.dat$concentration),
     xlab = "Time", ylab = "Log of concentration")
abline(lm.solution.log, col = "blue")
```



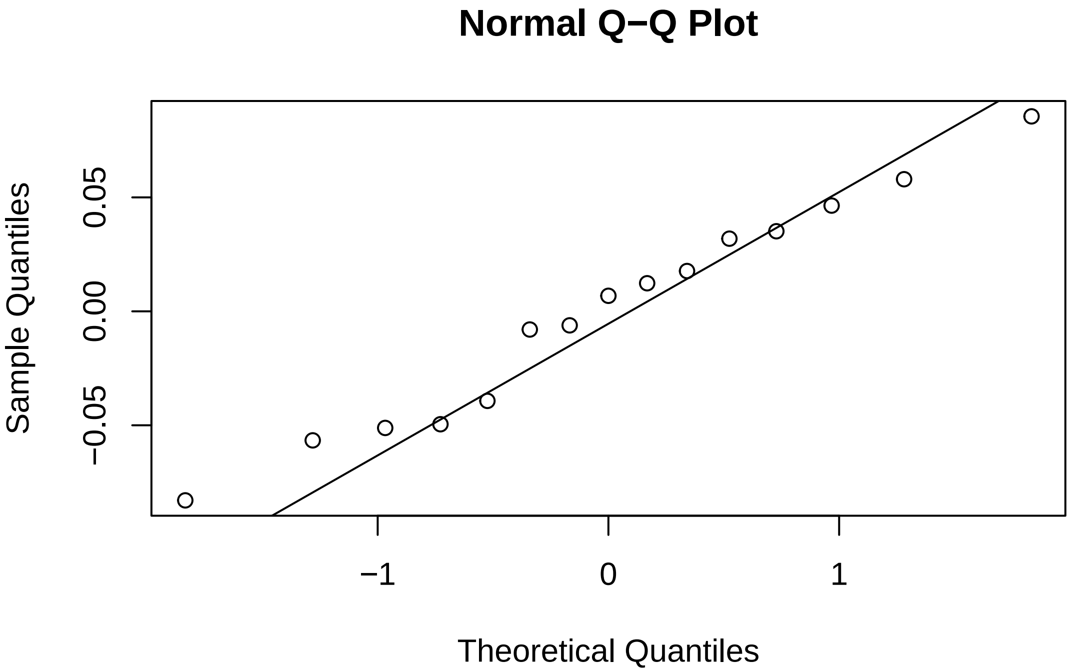
The regression line does appear to be a good fit to the transformed data.

Part (e)

```
solution.res.log <- lm.solution.log$residuals
solution.fit.log <- lm.solution.log$fitted.values

# Residual Plot
plot(solution.fit.log, solution.res.log, xlab = "Fitted values",
     ylab = "Residuals", main = "Residuals vs. Fitted Values")
abline(h=0)
```



```
# Normal Probability Plot
qqnorm(solution.res.log)
qqline(solution.res.log)
```



The residual plot seems to indicate that there are no major violations of the constant variance or linearity assumptions. The normal probability plot indicates that there appears to be no major violations of the normality assumption. Thus, we have evidence that the linear model is a good fit to the log-transformed data.

Part (f)

Our model is $\log_{10}(\hat{Y}_i) = 0.655 - 0.195 \cdot \text{time}_i$. Thus, to get everything back to the original units, we simply do:

$$10^{\log_{10}(\hat{Y}_i)} = 10^{0.655 - 0.195 \cdot \text{time}_i}$$
$$\hat{Y}_i = 4.518 \cdot 0.638^{time_i}$$

Problem 4.6

```
MM.dat <- read.table("./BookDataSets/Chapter  1 Data Sets/CH01PR27.txt", header = F)
names(MM.dat) <- c("MuscleMass", "Age")
```

Part (a)

To obtain 99 percent family confidence intervals for β_0 and β_1, $\alpha = 0.01$, therefore we need to construct $1 - \alpha/2 = 0.995 \times 100\% = 99$ percent intervals for each of the two parameters.

```
MM.fit <- lm(MuscleMass ~ Age, data=MM.dat)
confint(MM.fit, level=0.995)
```

```
##                  0.25 %     99.75 %
## (Intercept) 140.259608 172.4335205
## Age          -1.453227  -0.9267644
```

Part (b)

Opposite directions because b_0 and b_1 are negatively correlated.

Part (c)

No it does not because these values are not inside the 99 percent family confidence intervals.