STAT 516 hw 2

Solutions

Chp 8 Ex 3

The code below reads in the asphalt data set.

```
asphalt <- read.table(file = "Data Tables 4th edition/Chapter 8/datatab_8_24.prn",
                      header = TRUE)
head(asphalt)
```

```
  obs  x1   x2  x3  y1   y2
1   1 5.3 0.02  77  42 3.20
2   2 5.3 0.02  32 481 0.73
3   3 5.3 0.02   0 543 0.16
4   4 6.0 2.00  77 609 1.44
5   5 7.8 0.20  77 444 3.68
6   6 8.0 2.00 104 194 3.11
```

Stress at which a specimen fails

First we regress the stress at which a specimen failed (Y_1) on the predictor variables. We fit a multiple linear regression model with the `lm()` function.

```
lm_stress <- lm(y1 ~ x1 + x2 + x3, data = asphalt)
summary(lm_stress)
```

```
Call:
lm(formula = y1 ~ x1 + x2 + x3, data = asphalt)

Residuals:
```

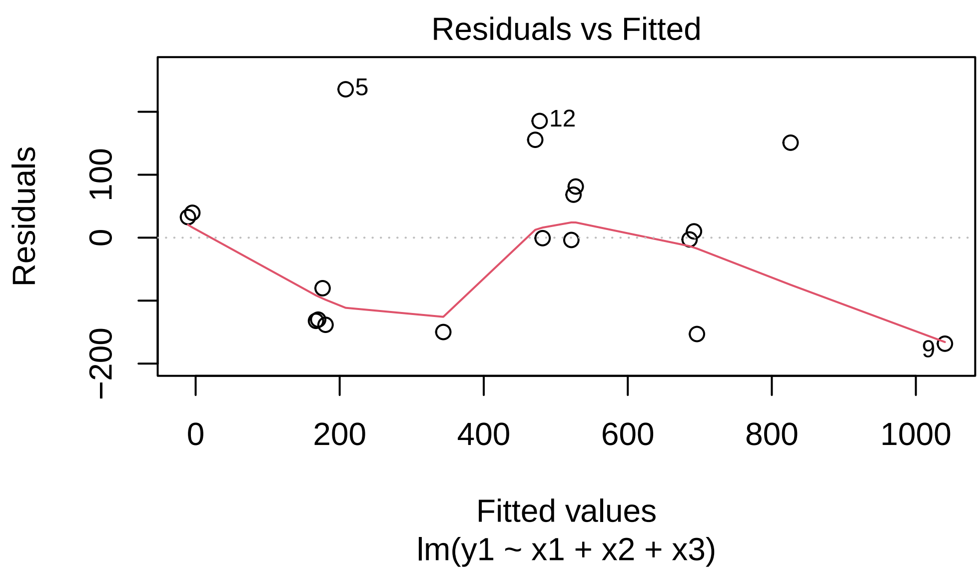
```
    Min       1Q   Median       3Q      Max 
-168.380 -131.124   -0.743   74.773  235.765 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 700.6180   125.8722   5.566 5.40e-05 ***
x1           -1.5257    13.0242  -0.117 0.908302    
x2          175.9839    35.6550   4.936 0.000179 ***
x3           -6.6971     0.8847  -7.570 1.69e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.9 on 15 degrees of freedom
Multiple R-squared:  0.8376,    Adjusted R-squared:  0.8051 
F-statistic: 25.79 on 3 and 15 DF,  p-value: 3.599e-06
```

Before interpreting the results, we check the residuals versus fitted values plot to see if there is any pattern in the residuals that would indicate nonlinearity in the relationship of the response to the predictors or nonconstant variance of the response given the predictors.

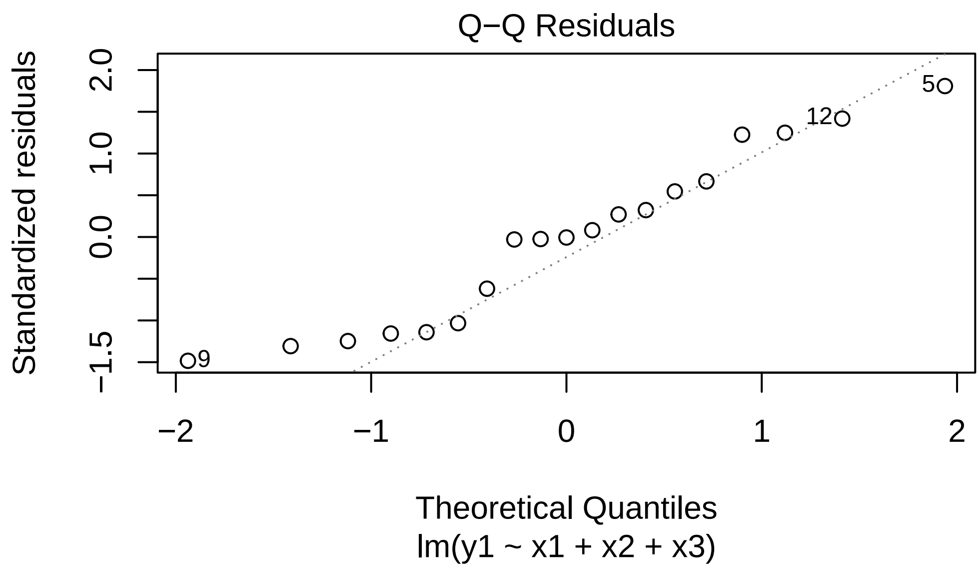
```
plot(lm_stress,which=1)
```



The red line which the `plot()` function draws through the points in the residuals versus fitted values plot suggests nonlinearity in the relationship between the response and the regressor variables. However, it is based on quite a small number of points, so the strength of the suggestion is quite small. If no red line were plotted, one probably not from this plot suspect nonlinearity. It is likely safe to assume the the relationship, if not exactly linear, is close enough to linear for the linear model to be useful.

We now check the Normal quantile-quantile plot of the residuals to see if we should assume Normality of the error terms in the multiple linear regression model.

```
plot(lm_stress,which = 2)
```



The Normal quantile-quantile plot indicates some departure from Normality in the lower tail of the distribution of the residuals (lower left part of the plot). Apart from this, the data points fall roughly along a straight line, and so it is likely safe to proceed under the assumption that the error terms have the Normal distribution.

Taking the assumptions of the multiple linear regression model to be satisfied, we may now interpret the output printed by the `summary()` function applied to the linear model object returned by the `lm()` function.

We see that the fitted model is

$$Y_1 = 700.62 + -1.53X_1 + 175.98X_2 + -6.7X_3,$$

according to which the stress at which a specimen fails (Y_1) is negatively affected by increases in the percent binder (X_1) and in the ambient temperature (X_3) and positively affected by the loading rate (X_2).

The p-values for testing H_0: $\beta_j = 0$ for $j = 1, 2, 3$, indicate that the estimated effect of percent binder (X_1) may be spurious—that is, it is not different enough from zero to be statistically significant, as its p-value is very large. The estimated effects of ambient temperature (X_2) and loading rate (X_3), however, do appear to reflect real effects, as the p-values are very small.

The code below prints confidence intervals for the coefficient values.

```
confint(lm_stress)
```

```
                 2.5 %     97.5 %
(Intercept) 432.327719 968.908377
x1          -29.286064  26.234706
x2           99.987067 251.980812
x3           -8.582839  -4.811437
```

We see that the 95% confidence intervals for the ambient temperature (X_2) and loading rate (X_3) coefficients do not contain zero, whereas that of percent binder (X_1) does contain zero; so it is plausible that percent binder (X_1) has no real linear relationship with the stress at which a specimen fails (Y_1).

Strain at which a specimen fails

Now we carry out a similar analysis with the strain at which a specimen failed (Y_2) as the response.

```
lm_strain <- lm(y2 ~ x1 + x2 + x3, data = asphalt)
summary(lm_strain)
```

```
Call:
lm(formula = y2 ~ x1 + x2 + x3, data = asphalt)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5466 -1.4827 -0.1190  0.6097  5.0135 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.61130    2.04575  -2.743 0.015100 *  
x1           0.66754    0.21168   3.154 0.006558 ** 
x2          -1.23535    0.57949  -2.132 0.049966 *  
x3           0.07319    0.01438   5.090 0.000133 ***
---
```

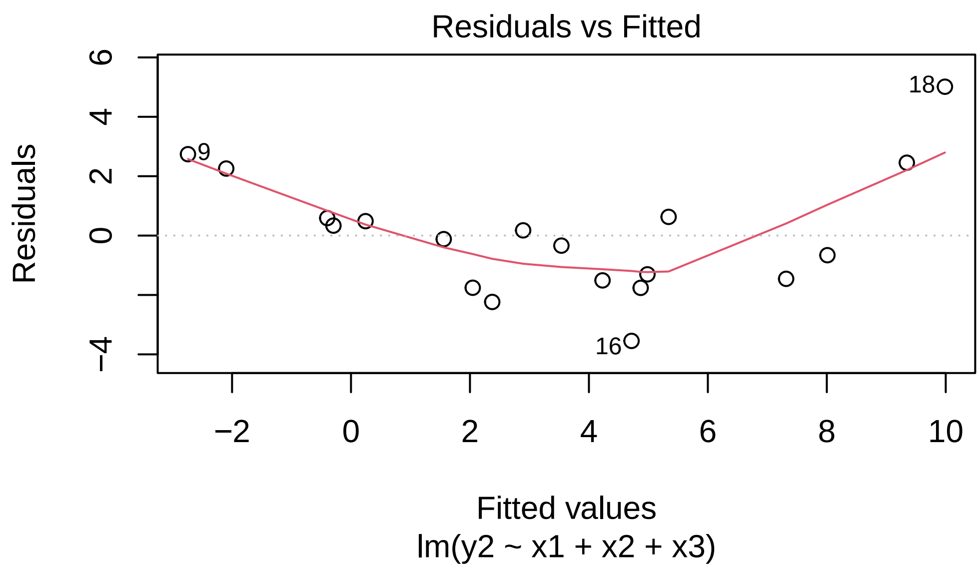
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.241 on 15 degrees of freedom
Multiple R-squared:  0.7601,    Adjusted R-squared:  0.7121
F-statistic: 15.84 on 3 and 15 DF,  p-value: 6.447e-05
```

Before we interpret the results, we check whether the multiple linear regression assumptions are satisfied.

First we look at the residuals versus fitted values plot:

```
plot(lm_strain,which=1)
```

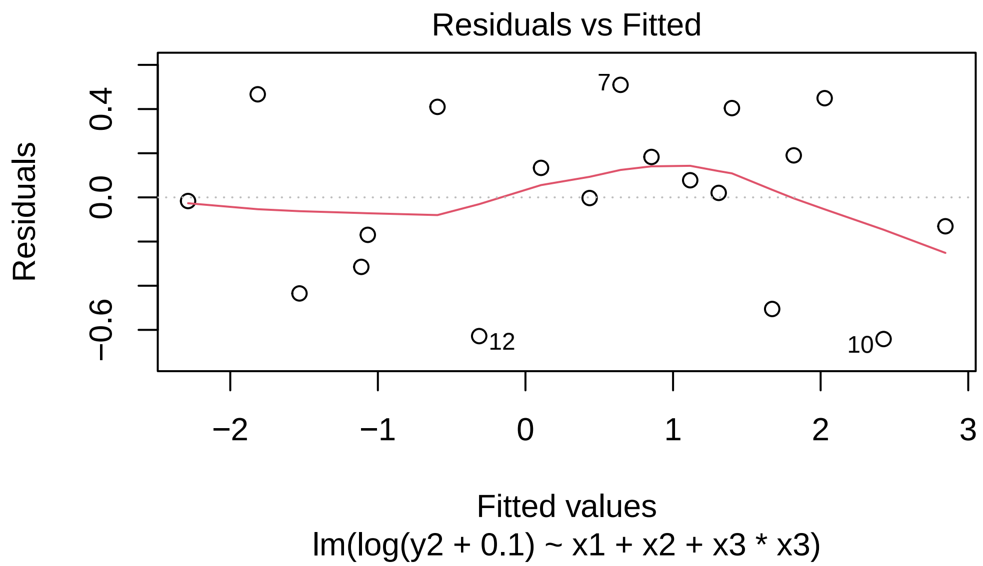


In this plot we see a pretty clear indication of nonlinearity in the relationship between the response and the covariates. Even if the red line were removed, the 'swoosh' pattern in the points would still be apparent.

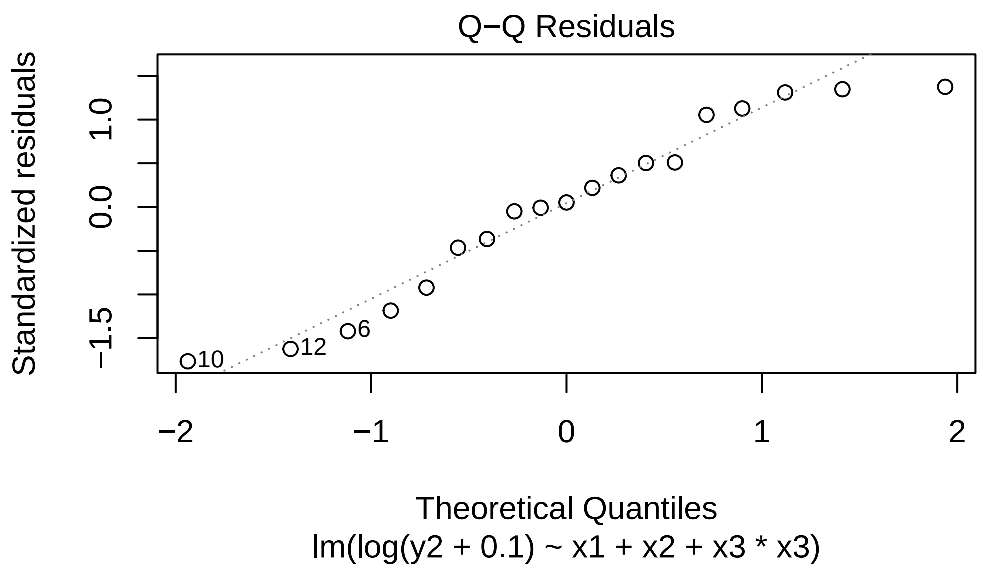
Unless we transform the data, the analysis should stop here, because the assumptions of the multiple linear regression model are not satisfied.

We cannot take the natural log of Y_2, because one of the values is zero. In this case, one can try adding a small constant to all the values and then taking the log. Let's consider the transformed response $\log(Y_2 + 0.1)$.

```
lm_logstrain <- lm(log(y2+.1) ~ x1 + x2 + x3*x3, data = asphalt)
plot(lm_logstrain,which = 1)
```




```r
plot(lm_logstrain,which = 2)
```



Now the residuals versus fitted values plot and the Normal quantile-quantile plot suggest that the multiple linear regression assumptions are satisfied under the transformed response $\log(Y_2 + 0.1)$.

The fitted model is

$$\log(Y_2 + 0.1) = -2.361 + 0.105X_1 + -0.381X_2 + 0.038X_3.$$

Below is a summary of the linear model fit:

```r
summary(lm_logstrain)
```

```

Call:
lm(formula = log(y2 + 0.1) ~ x1 + x2 + x3 * x3, data = asphalt)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.64147 -0.24218  0.02024  0.29730  0.50954 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.36132    0.37276  -6.335 1.34e-05 ***
x1           0.10472    0.03857   2.715  0.01597 *  
x2          -0.38124    0.10559  -3.611  0.00257 ** 
x3           0.03805    0.00262  14.525 3.05e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4083 on 15 degrees of freedom
Multiple R-squared:  0.9423,	Adjusted R-squared:  0.9308 
F-statistic: 81.66 on 3 and 15 DF,  p-value: 1.615e-09
```

From the summary we can see that all three of the covariates appear to have significant effects on the transformed response, as the p-values are all quite small.

An interpretation, for example, of the estimated coefficient on X_1, is that for an increase in X_1, the percent binder, of one unit, the strain at which a specimen fails increases by 10.5 percent (ignoring the small constant 0.1 that we added to all the response values before the log transformation).

Chp 8 Ex 5

We first read in the data:

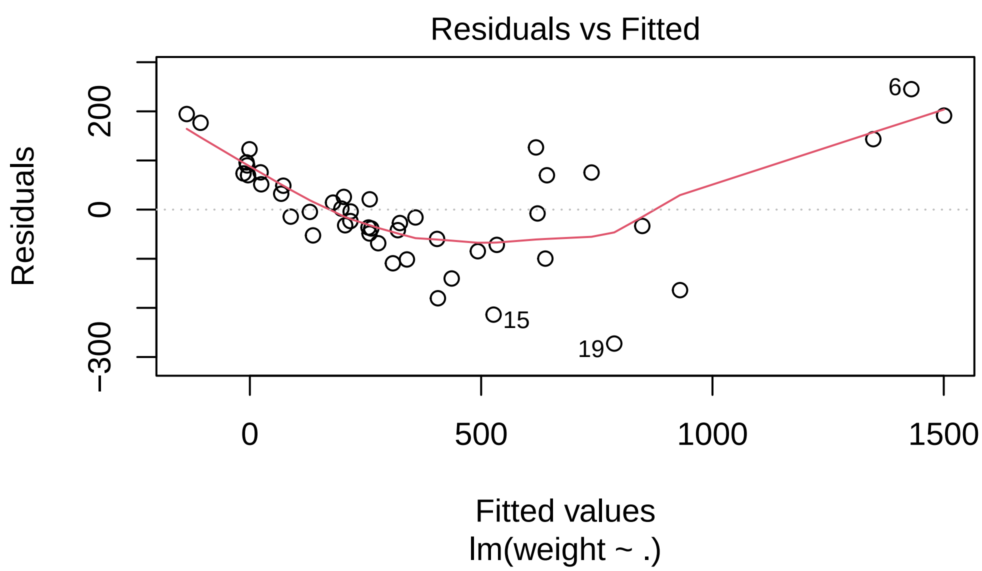
```r
tree <- read.table(file = "Data Tables 4th edition/Chapter 8/datatab_8_26.prn",
                   header = TRUE)
head(tree)
```

```
  obs  dbh height age  grav weight
1   1  5.7     34  10 0.409    174
2   2  8.1     68  17 0.501    745
3   3  8.3     70  17 0.445    814
4   4  7.0     54  17 0.442    408
5   5  6.2     37  12 0.353    226
6   6 11.4     79  27 0.429   1675
```

a)

Now we fit a multiple linear regression model with the variable `weight` as the response and make the residuals versus fitted values plot.

```
lm_tree <- lm(weight ~ ., data = tree)
plot(lm_tree,which = 1)
```



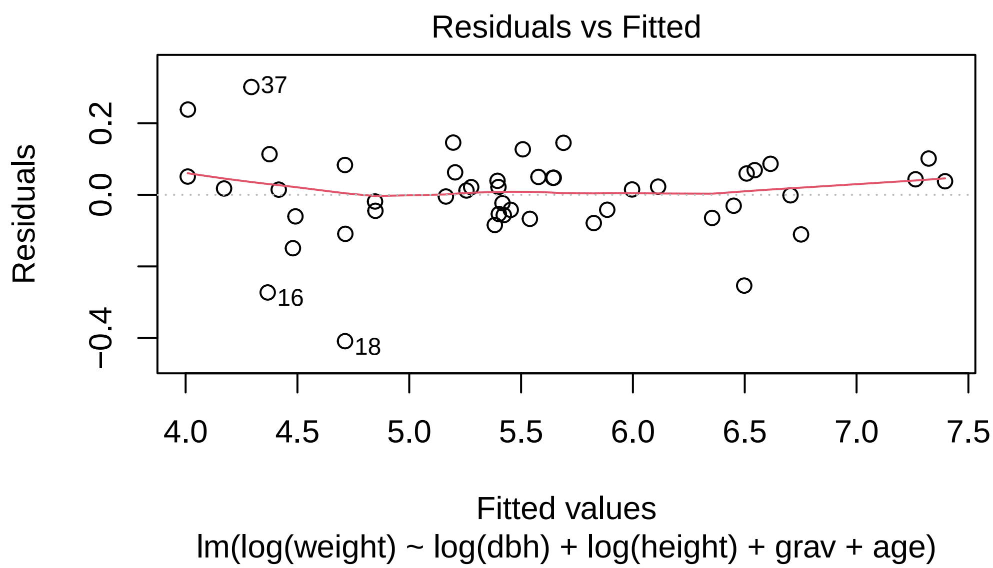
From the residuals vs fitted values plot, the relationship between the response and the covariates appears to be nonlinear. So the fitted model is not useful.

b)

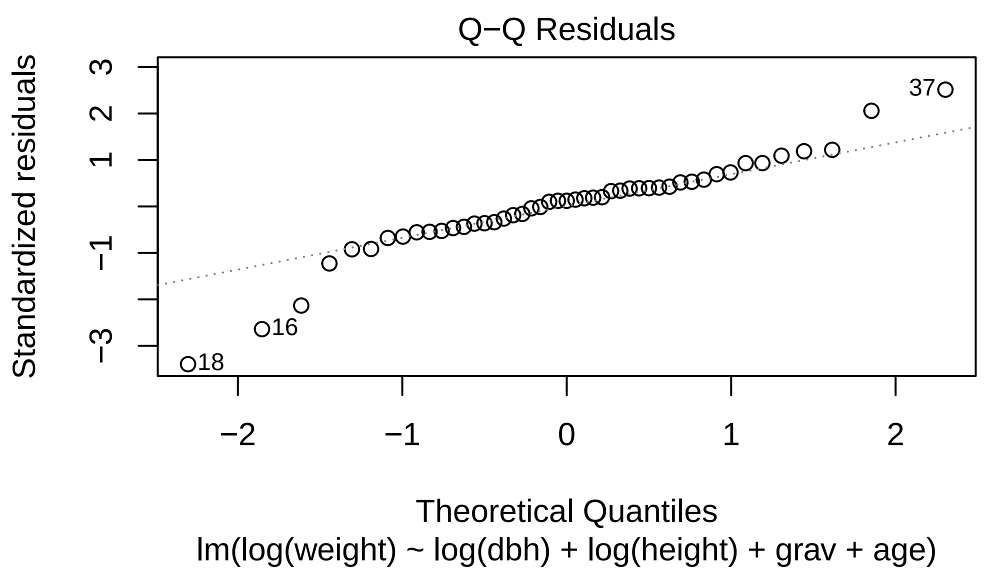
Since it is natural to assume the weight is equal to something like the height times the diameter (the volume), that is weight $\approx$ height $\times$ diameter, then by the rules of logarithms log(weight) $\approx$

log(height) + log(diameter). This suggests fitting a linear model after log-transforming these three variables.

```
lm_logtree <- lm(log(weight) ~ log(dbh) + log(height) + grav + age, data = tree)
plot(lm_logtree,which = 1)
```




```
plot(lm_logtree,which = 2)
```



The residuals versus fitted values plot indicates that the linear model is a good fit to the data, and the Normal quantile-quantile plot, in spite of showing a few low-outlying residuals, suggests that it is likely safe to assume that the error terms follow a Normal distribution. Actually, since the sample size is rather large ($n \geq 30$), it is likely safe to assume that the least squares estimators of the regression coefficients have approximately a Normal distribution even if the error terms do not.

From here, we note that the fitted model is

$$\log(\text{weight}) = -1.984 + 2.156 \log(\text{dbh}) + 0.968 \log(\text{height}) + 0.176\text{grav} + -0.009\text{age},$$

which we can see from the summary:

```r
summary(lm_logtree)
```

```

Call:
lm(formula = log(weight) ~ log(dbh) + log(height) + grav + age,
    data = tree)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40874 -0.05508  0.01500  0.05526  0.30120

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.983517   0.418596  -4.738 2.48e-05 ***
log(dbh)     2.156404   0.116623  18.490  < 2e-16 ***
log(height)  0.968157   0.162623   5.953 4.64e-07 ***
grav         0.175618   0.608581   0.289    0.774
age         -0.009175   0.004463  -2.056    0.046 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1257 on 42 degrees of freedom
Multiple R-squared:  0.9822,	Adjusted R-squared:  0.9805
F-statistic:   579 on 4 and 42 DF,  p-value: < 2.2e-16
```

From the summary we can also see that the effects of log(dbh) and log(height) have very small p-values, so their effect on the weight is highly significant. The covariate grav does not appear to have a significant effect, and the covariate age has an effect which is only barely significant if one compares it to a 0.05 significance level.

Chp 8 Ex 7

Here we read the data into R:

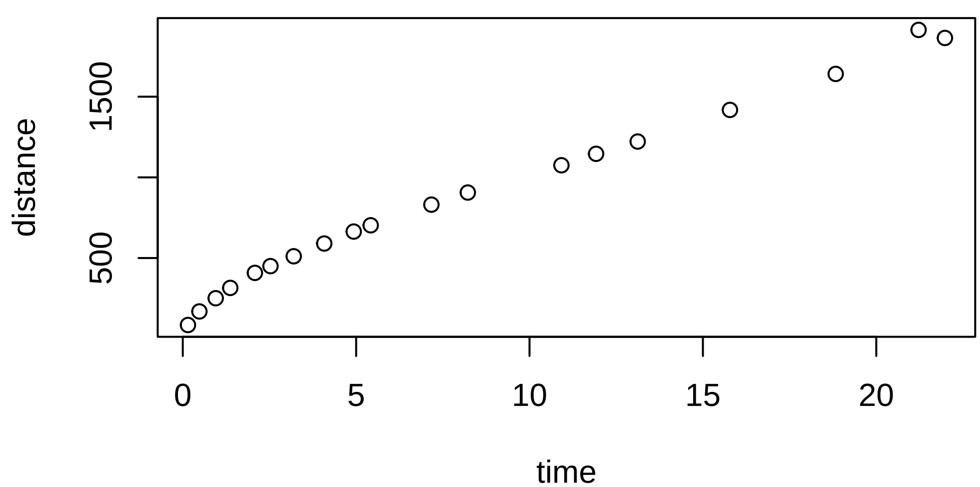
```
irrigation <- read.table(file = "Data Tables 4th edition/Chapter 8/datatab_8_28.prn",
                         header = TRUE)
head(irrigation)
```

```
  obs distance time
1   1       85 0.15
2   2      169 0.48
3   3      251 0.95
4   4      315 1.37
5   5      408 2.08
6   6      450 2.53
```

a)

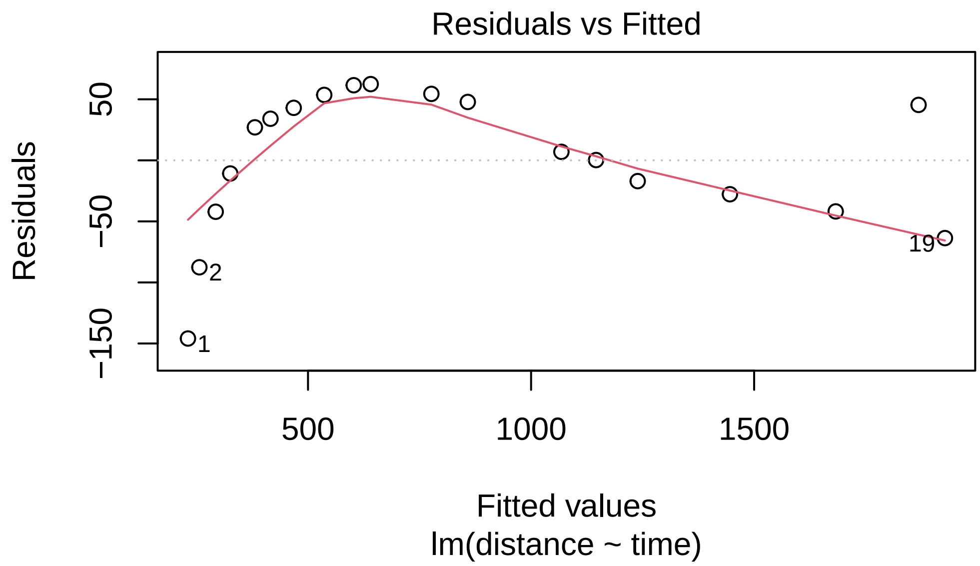
The code below makes a scatterplot of the distance versus the time variable.

```
plot(distance ~ time, data = irrigation)
```



The relationship appears nonlinear. The following code fits a simple linear regression model and produces the residuals versus fitted values plot.

```
lm_out <- lm(distance ~ time, data = irrigation)
plot(lm_out, which = 1)
```

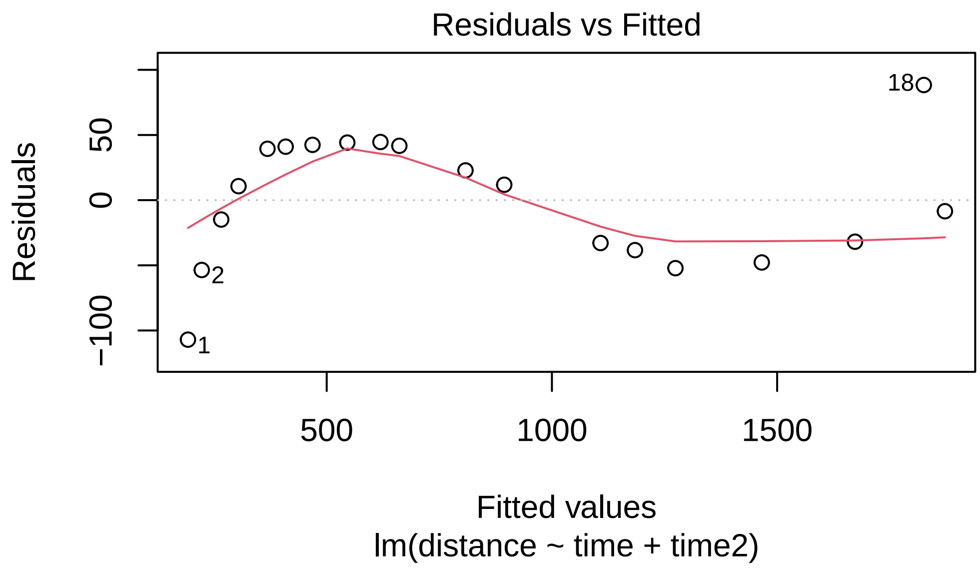


The nonlinearity in the relationship between the distance and time variables is much more apparent in the residuals versus fitted values plot.

b), c)

We now fit a model which includes time and the square of time.

```
irrigation$time2 <- irrigation$time^2
lm2_out <- lm(distance ~ time + time2, data = irrigation )
plot(lm2_out,which = 1)
```



This model is still a poor fit!