Examples of Applying Figure Sense in Model Building

Example 1: **An experiment was conducted in which the response was the survival time in units of 10 hours of animals that were given one of three poisons and then administered one of 4 treatments to counteract the poison. The experiment was part of an investigation to combat the effects of certain toxic agents. The data from the experiment are in the following table.**

Survival Times

	Treatment A	Treatment B	Treatment C	Treatment D
Poison 1	.31 .45 .46 .43	.82 1.10 .88 .72	.43 .45 .63 .76	.45 .71 .66 .62
Poison 2	.36 .29 .40 .23	.92 .61 .49 1.24	.44 .34 .31 .40	.56 1.02 .71 .38
Poison 3	.22 .21 .18 .23	.30 .37 .38 .29	.23 .25 .24 .22	.30 .36 .31 .33

Evaluate the impact of the poison and treatments of the survival times of the animals.

Figure Sense Habit: Define the Problem:

What do I know? or What information do I have to work with?

What do I want to accomplish?

What steps to I need to take to get from what I know to what I want to accomplish?

Step 1: What do I know?

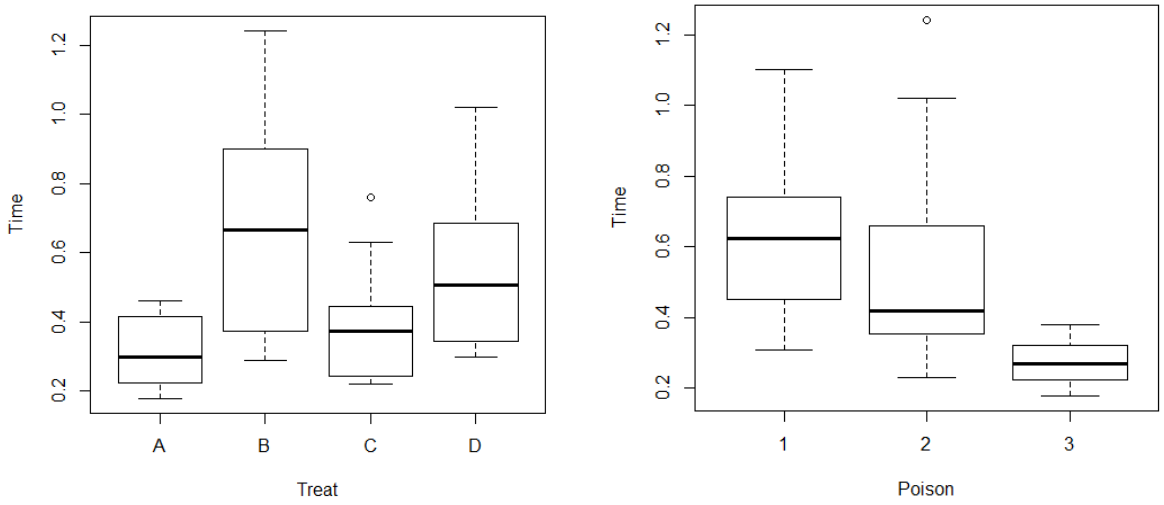
This is a designed experiment with two factors: Poison at 3 different levels and Treatment at 4 different levels. This was replicated four times for each combination of Poison and Treatment.

A standard way to analyze the data is a two way Analysis of Variance (ANOVA). This analysis will determine if the mean survival time is affected by one of both of the Poison or Treatment.

A two way ANOVA is based upon a set of assumptions being approximately true.

Plots of the data:

The following are Box plots of the survival time for the four different treatments and the three different poisons.



The medians of the survival time changes with both the treatment and the poison, thus these variables should be good inputs to a model to predict survival times.

The variances of the survival time distributions do not seem to be the same for the different treatments or for the different poisons.

Figure Sense Habit: Evaluate the assumptions that are necessary to solve the problem.

What assumptions do I need to make to solve this problem?

Are these assumptions approximately true for this problem?

If the assumptions are not true for this problem, what needs to be changed?

What assumptions are necessary to use a two way ANOVA analysis?

1) The mean responses are a function of the values of the two factors.

2) The observed values are equal to the mean responses plus an error term.

3) The error terms have an approximate Normal distribution with mean equal to 0 and a constant variance. The errors for any two observed values are statistically independent.

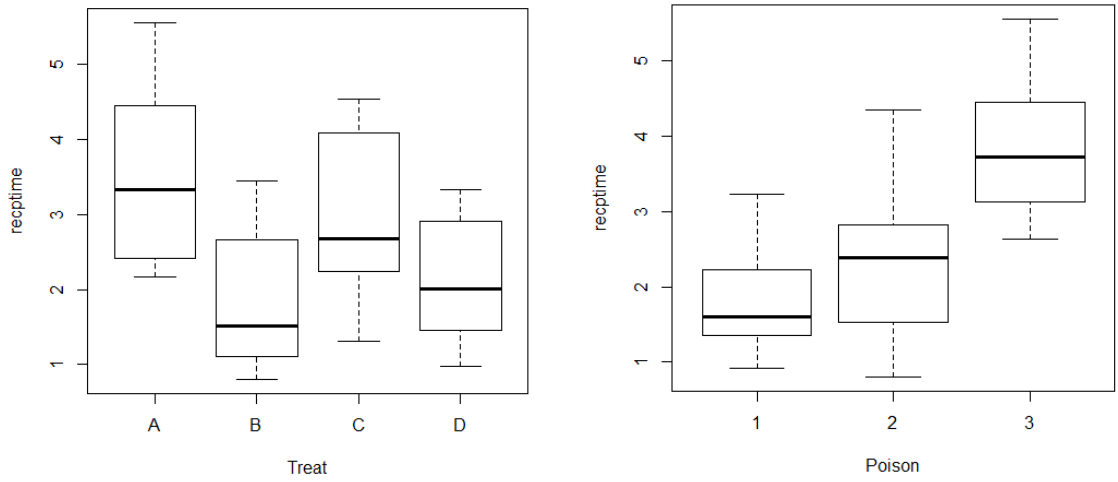
Are these assumptions approximately true for this data?

No, the variances for the different treatment groups do not appear to be the same and the variances for the different poisons do not appear to be the same.

What needs to be changed do the assumptions are approximately correct?

Since the groups with the larger variances are also the groups with the larger medians. It may be that a nonlinear transformation of the response variable, survival time, will correct the problem with the non-constant variance.

If we transform the response variable to recptime = 1/Time, then the plots for the transformed variable follow.



The medians of the reciprocal of survival time changes with both the treatment and the poison, thus these variables should be good inputs to a model to predict recptime.

The variances of the reciprocal of survival time seem relatively consistant for the different treatments or for the different poisons.

Step 2: What do I want to accomplish?

Do the different treatments and the different poisons affect the survival times of the animals.

Step 3: What steps do I need to take to get from what I know to what I what to accomplish?

Estimate a two way ANOVA model and perform diagnostic checks on the residuals to determine if the model assumptions are approximately correct for the response variable recptime. If there are no problems with the residual checks the 2 way ANOVA model can be used to answer the question of interest.

Figure Sense Habit: Look for unusual outcomes or exceptions.

Before solving the problem ask: What do I expect the answer to be?

After solving the problem ask: Is the answer consistent with what I expected?

What do I expect the answer to be?

If the model assumptions are approximately correct for the response variable recptime, then plots of the residuals should exhibit the following characteristics:

a) The plot of the residuals versus the fitted values should look there is no relationship between the residuals and the fitted values and the variance of the residuals should be constant as the fitted values change.

b) The plot of the residuals versus the treatments should have no relationship between the residuals and the treatments and the variance of the residuals should be constant as the treatment values change.

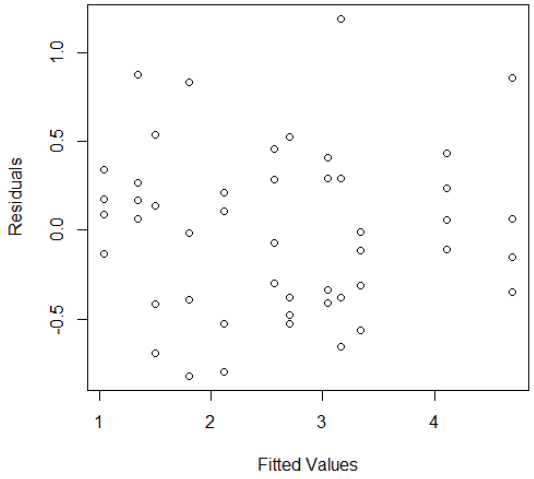
c) The plot of the residuals versus the poisons should have no relationship between the residuals and the poisons and the variance of the residuals should be constant as the poison values change.

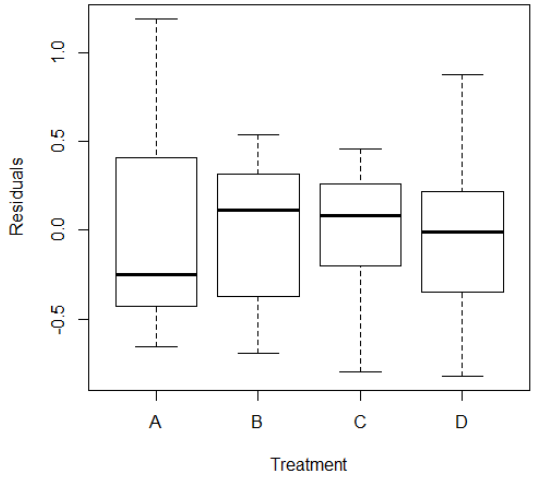
d) The residuals should be approximately Normally distributed.

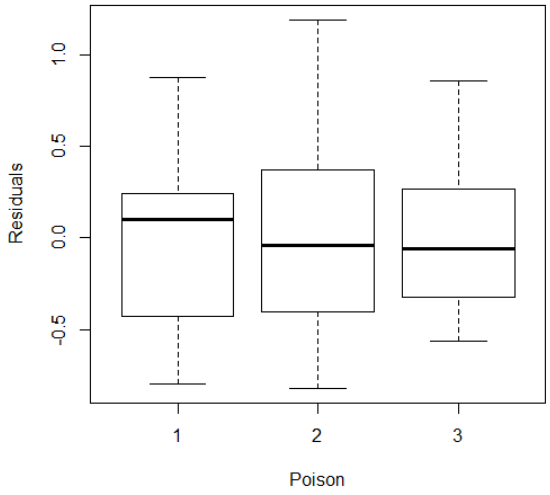
Solve the problem:

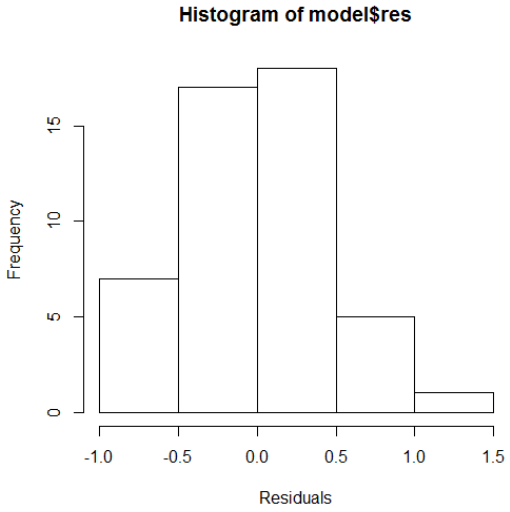
The output for a 2 way ANOVA with the response recptime and the residual plots for this model follow:

```
Analysis of Variance Table

Response: recptime
          Df Sum Sq Mean Sq F value    Pr(>F)
Treat      3 20.290  6.7632  28.405 3.404e-10 ***
Poison     2 35.017 17.5084  73.533 1.903e-14 ***
Residuals 42 10.000  0.2381
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```










```
        Shapiro-Wilk normality test

data:  model$res
W = 0.9781, p-value = 0.5023
```

Are the residuals consistent with what I expected?

a) In the plot of the residuals versus the fitted values, the residuals do not suggest any relationship with the fitted values, they have a mean of 0 and a constant variance as the fitted values change. This is consistent with what is expected.

b) In the plot of the residuals versus the Treatment, the mean is near zero and the variation is consistent for each of the 4 treatments. This is consistent with what is expected.

c) In the plot of the residuals versus the Poison, the mean is near zero and the variation is consistent for each of the 3 poisons. This is consistent with what is expected.

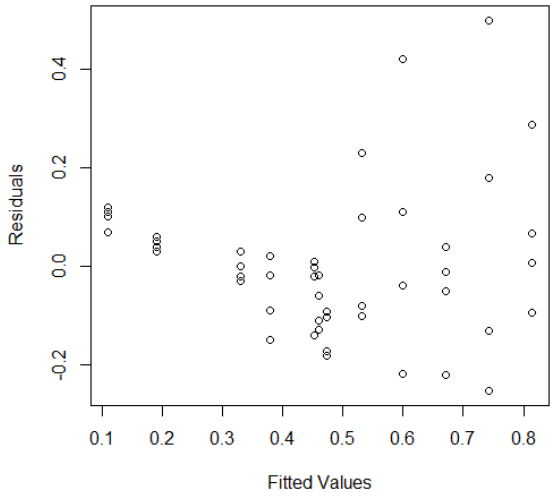
d) The histogram is consistent with the residuals having a Normal distribution and the null hypothesis that the residuals follow a Normal distribution cannot be rejected (in the Shapiro-Wilk test). This is consistent with what is expected.

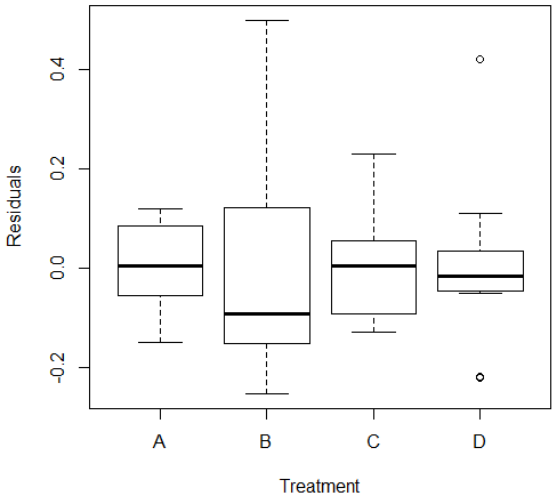
Since the residuals behave as expected if the model assumptions are correct, then the model can be used to answer the question of whether the treatments or poisons impact the survival time. For this document this analysis will be omitted.

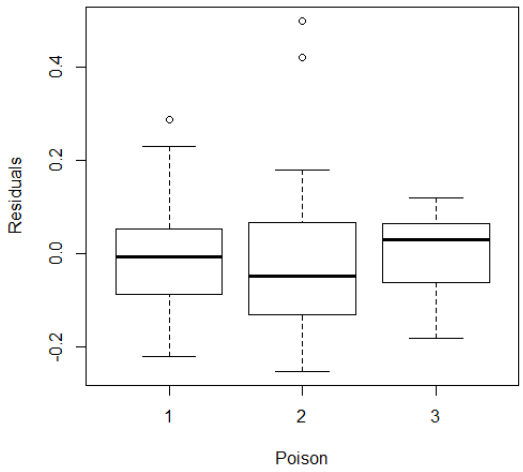
Alternative Analysis of this data:

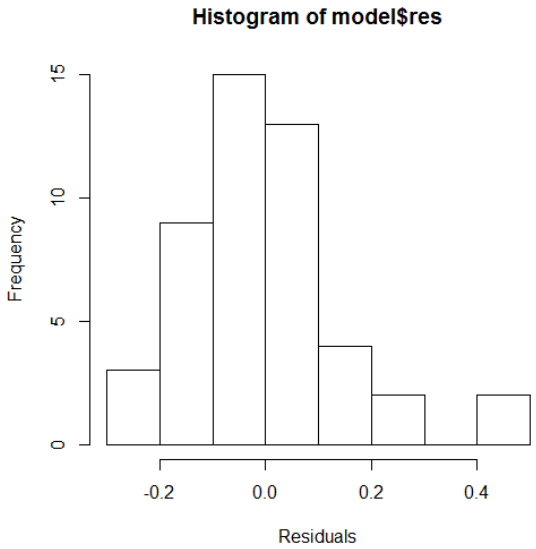
If an analyst did not recognize that it was necessary to transform the response to recptime and instead estimated a 2 way ANOVA model using the response the survival time, then the statistical output and the plots of the residuals for this analysis follow:

```
Analysis of Variance Table

Response: Time
          Df  Sum Sq Mean Sq F value    Pr(>F)
Treat      3 0.91776 0.30592  12.292 6.599e-06 ***
Poison     2 1.03563 0.51781  20.805 5.257e-07 ***
Residuals 42 1.04531 0.02489
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```









```
        Shapiro-Wilk normality test

data:  model$res
W = 0.92211, p-value = 0.003532
```

Are the residuals consistent with what I expected?

a) In the plot of the residuals versus the fitted values, there is a curved pattern in the plot and the variation of the residuals gets larger as the fitted values increase. This is not consistent with what is expected. This "cornucopia" pattern suggests the need to transform the response variable.

b) In the plot of the residuals versus the Treatment, the variation is not consistent between the four treatments. This is not consistent with what is expected.

c) In the plot of the residuals versus the Poison, the mean is near zero and the variation is consistent for each of the 3 poisons. This is consistent with what is expected.

d) The histogram is slightly skewed to the right, this is not consistent with a Normal distribution and the null hypothesis that the residuals follow a Normal distribution is rejected (in the Shapiro-Wilk test). This is not consistent with what is expected.

In the alternative analysis, since the characteristics of the residuals is not what we would expect if the assumptions for the 2 way ANOVA model are true, something needs to be done to change the model. Fortunately, the plot of the residuals versus the fitted values suggests that the response variable should be transformed. This would lead to using the reciprocal of the survival time as the new response variable.

Example 2: A data set contains the variables Brain = the brain size and Body = the corresponding body size for 96 animals. Analyze this data to determine a model that relates the variable Body to the variable Brain.

Figure Sense Habit: Define the Problem:

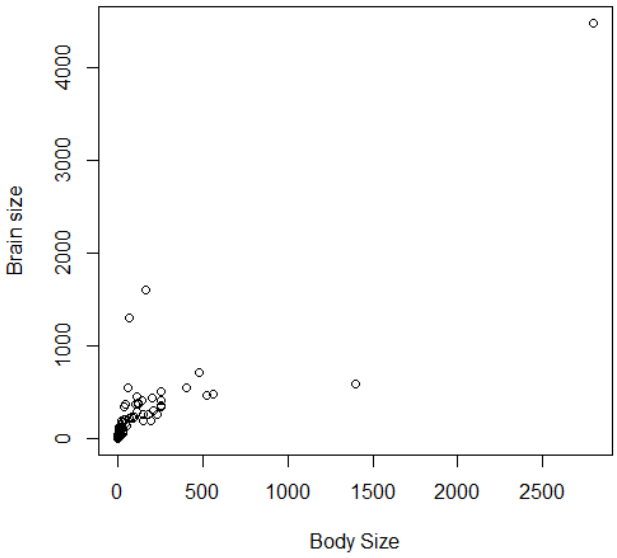
What do I know? or What information do I have to work with?

What do I want to accomplish?

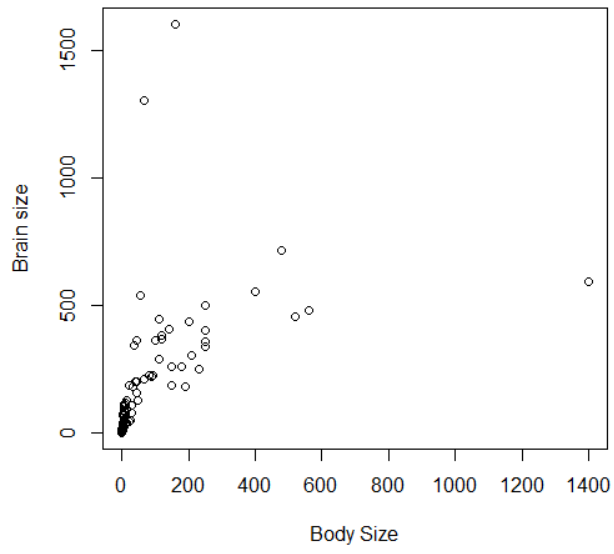
What steps to I need to take to get from what I know to what I want to accomplish?

Step 1: What do I know?

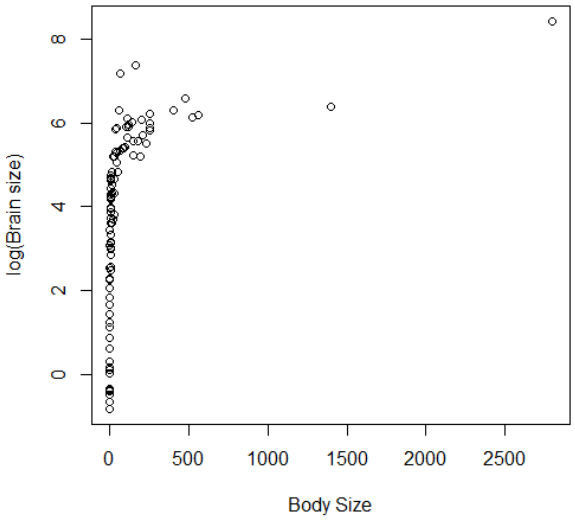
The data Brain and Body are available, to understand if there is a relationship between these two variables, it is a good idea to plot these variables. A scatter plot of the variable Brain versus the variable Body follows:



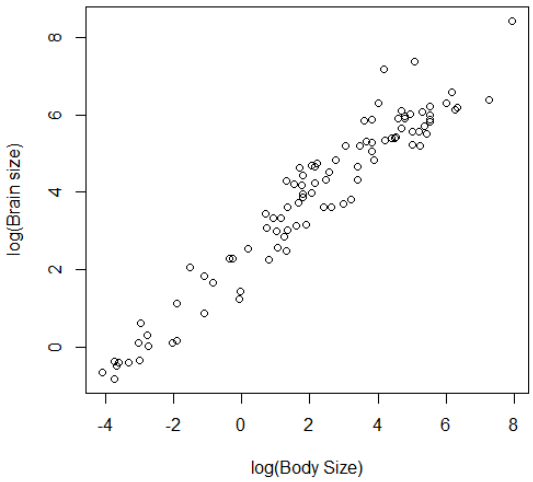
There is one animal, the African elephant, with a brain and body size much larger than the other animals. It may be helpful to redo this plot after removing the elephant since in the plot above the other animals are in a small portion of the plot.



Both of these plots suggest that there is a nonlinear relationship between the variables Brain and Body and suggest that the variance is not constant. These suggest the need to transform the response variable Brain. The following is a plot of log(Brain) versus Body.



In the above plot, there is a curved relationship that is monotonically increasing. It may be helpful to also transform the variable Body in the hope that the relationship will then be more linear. The following is a plot of the variable log(Brain) versus log(Body).



This plot suggests that there is a linear relationship between log(Brain) and log(Body) and that the variation around the line is relatively constant. Thus, it appears that a linear regression model with response variable log(Brain) and input variable log(Body) is appropriate.

Step 2: What do I want to accomplish?

Develop a model to relates the variable Body to the variable Brain.

Step 3: What needs to be done to get from what I know to what I want to accomplish?

Estimate a simple linear regression model with dependent variable log(Brian) and independent variable log(Body). Check the characteristics of the residuals to determine if they are consistent with the assumptions for a linear regression model.

Figure Sense Habit: Look for unusual outcomes or exceptions.

Before solving the problem ask: What do I expect the answer to be?

After solving the problem ask: Is the answer consistent with what I expected?

What do I expect the answer to be?

If the model assumptions are approximately correct for the response variable log(Brain), then plots of the residuals should exhibit the following:

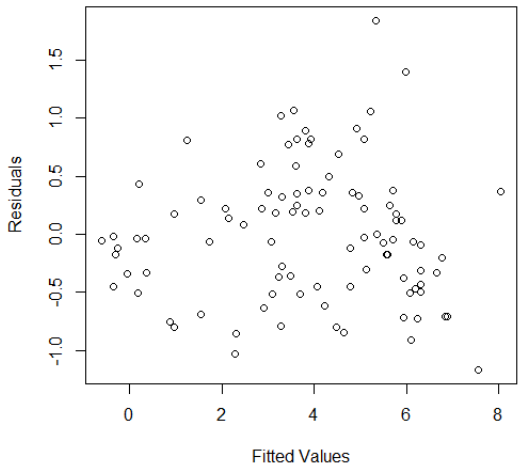
a) The plot of the residuals versus the fitted values should look there is no relationship between the residuals and the fitted values and the variance of the residuals should be constant as the fitted values change.

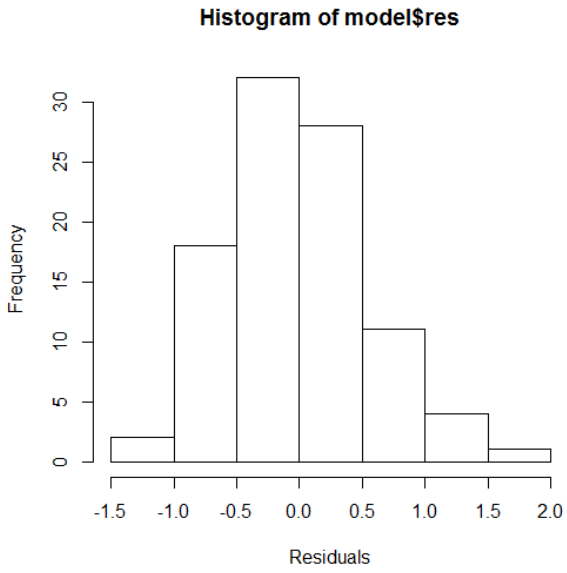
b) The residuals should be approximately Normally distributed.

Solve the problem:

The output from fitting a regression model with dependent variable log(Brain) and independent variable log(Body) and the plots of the residuals follow:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.33235    0.07325   31.84   <2e-16 ***
log(Body)    0.71919    0.02037   35.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5781 on 94 degrees of freedom
Multiple R-squared:  0.9299,Adjusted R-squared:  0.9291
F-statistic:  1246 on 1 and 94 DF,  p-value: < 2.2e-16
```






```
        Shapiro-Wilk normality test

data:  model$res
W = 0.98057, p-value = 0.1659
```

Are the residuals consistent with what I expected?

a) The plot of the residuals versus the fitted values does not show any relationship between the residuals and the fitted values and the variation of the residuals is fairly consistent as the fitted values change. This is consistent with what was expected.

b) The histogram is consistent with the residuals following a Normal distribution. The null hypothesis that the residuals follow a Normal distribution cannot be rejected (in the Shapiro-Wilk test). This is consistent with what was expected.

A regression model that describes the relationship between the size of the brain and the size of the body for the animals in this data set is:

$$\log(Brain)_i = 2.33 + .719 * \log(Body)_i + e_i$$

The error term e_i has a Normal distribution with mean = 0 and standard deviation = .5781.

Example 3: Old faithful geyser in Yellowstone National Park, Wyoming, derives its name from the regularity and beauty of its eruptions. As they do with most geysers in the park, rangers post the predicted times of eruptions on signs nearby, and people gather to witness the show. R. A. Hutchinson, a park geologist, collected measurements on the eruption durations (X, in minutes) and the subsequent intervals before the next eruption(Y, in minutes) over an 8 day period. Use this data

to develop a model to describe the relationship between the variable interval and the variable duration.

Figure Sense Habit: Define the Problem:

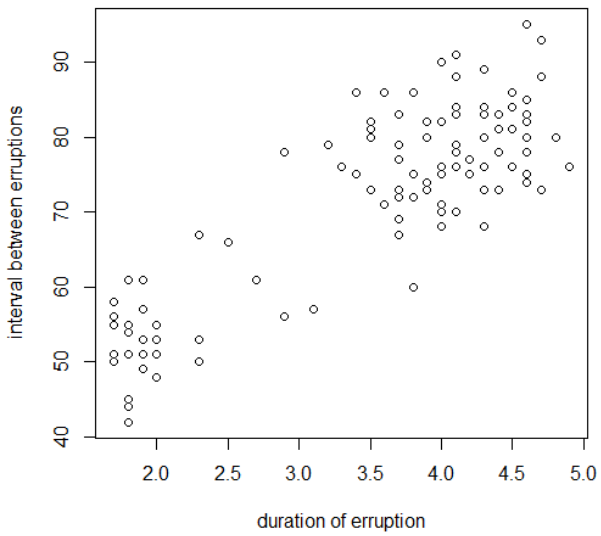
What do I know? or What information do I have to work with?

What do I want to accomplish?

What steps to I need to take to get from what I know to what I want to accomplish?

Step 1: What do I know?

The data interval and duration are available, to understand if there is a relationship between these two variables, it is a good idea to plot these variables. A scatter plot of the variable interval versus the variable duration follows:



The plot suggests that there is a linear relationship between the variable interval and the variable duration and that the variation around the line is consistent.

Step 2: What do I want to accomplish?

Develop a model that describes the relationship between the dependent variable interval and the independent variable duration.

Step 3: What needs to be done to get from what I know to what I want to accomplish?

Estimate a simple linear regression model with dependent variable interval and independent variable duration. Check the characteristics of the residuals to determine if they are consistent with the assumptions for a linear regression model.

Figure Sense Habit: Look for unusual outcomes or exceptions.

Before solving the problem ask: What do I expect the answer to be?

After solving the problem ask: Is the answer consistent with what I expected?

What do I expect the answer to be?

If the model assumptions are approximately correct, then plots of the residuals should exhibit the following:

a) The plot of the residuals versus the fitted values should look there is no relationship between the residuals and the fitted values and the variance of the residuals should be constant as the fitted values change.

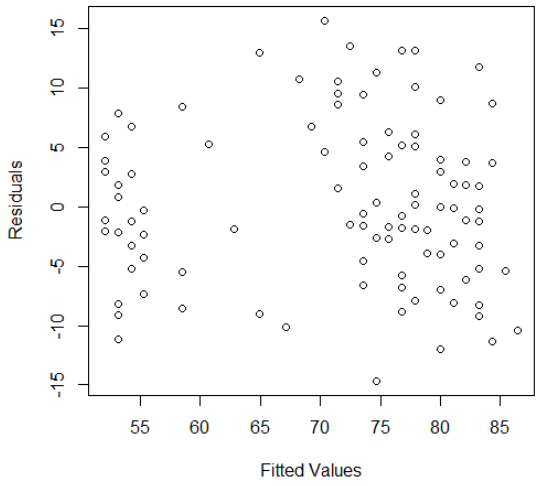
b) The residuals should be approximately Normally distributed.

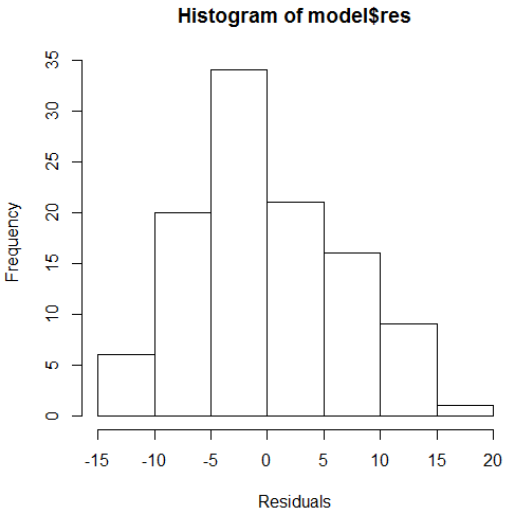
c) Since the data is a time series (the data was gathered sequentially in time) there should not be any large autocorrelations between the residuals. This is because the assumption is that the errors in the regression model are independent.

Solve the problem:

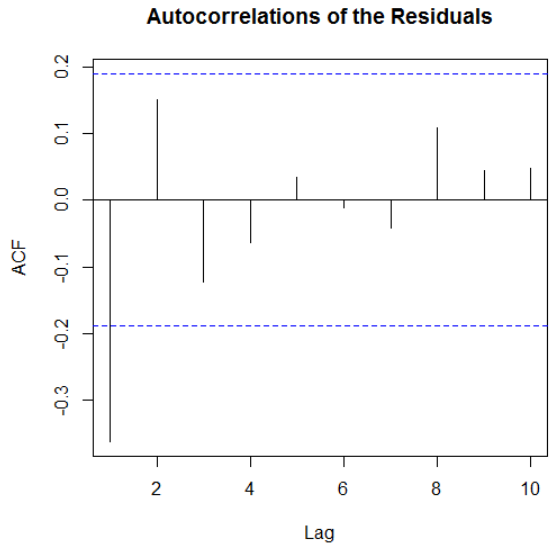
The regression output with interval as the dependent variable and duration as the independent variable and the plots of the residuals follow:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.8282     2.2618   14.96   <2e-16 ***
duration     10.7410     0.6263   17.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.683 on 105 degrees of freedom
Multiple R-squared:  0.7369,Adjusted R-squared:  0.7344
F-statistic: 294.1 on 1 and 105 DF,  p-value: < 2.2e-16
```






```
        Shapiro-Wilk normality test

data:  model$res
W = 0.98389, p-value = 0.2231
```



Are the residuals consistent with what I expected?

a) The plot of the residuals versus the fitted values does not show any relationship between the residuals and the fitted values and the variation of the residuals is fairly consistent as the fitted values change. This is consistent with what was expected.

b) The histogram is consistent with the residuals following a Normal distribution. The null hypothesis that the residuals follow a Normal distribution cannot be rejected (in the Shapiro-Wilk test). This is consistent with what was expected.

c) The lag 1 autocorrelation is substantially different from the value 0 which is what we expected if the model assumptions are correct. This suggests that there is a problem with this model.

What needs to be changed to address the inconsistency?

The regression model is:

$$interval_t = \beta_0 + \beta_1 duration_t + e_t$$

where the error terms, e_i, are all independent.

Because the residuals exhibit significant autocorrelations, the model can be modified as follows:

$$interval_t = \beta_0 + \beta_1 duration_t + N_t$$

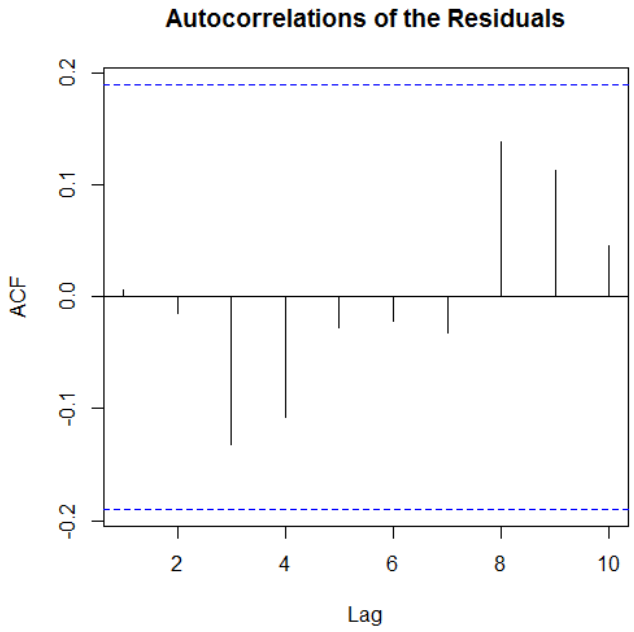
where the error term N_i follows the first order autoregressive model: $N_t = \phi N_{t-1} + e_t$

The output from this model follows:

```
Call:
arimax(x = interval, order = c(1, 0, 0), xreg = data.frame(duration),
method = "ML")

Coefficients:
          ar1  intercept  duration
      -0.4465    38.9879    9.2497
s.e.   0.0986     2.7831    0.7966

sigma^2 estimated as 36.82:  log likelihood = -344.86,  aic = 695.72
```

The estimated parameters and their standard errors are $\hat{\beta}_0$ = 38.988 SE($\hat{\beta}_0$) = 2.783; $\hat{\beta}_1$ = 9.250 SE($\hat{\beta}_1$) = .797; $\hat{\phi}$ = -.447 SE($\hat{\phi}$) = .099; and $\hat{\sigma}^2$ = 36.82 so that $\hat{\sigma}$ = 6.068

The plot of the autocorrelations of the residuals below shows that all the autocorrelations are sufficiently small to be statistically insignificant.



This model fits the data better than the simple linear regression model.