Design of Experiment for STAT112

Charles Fleming

March 24, 2018

Contents

Chapter 1

Theory

Let us start with the two parameter fixed effects linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \tag{1.1}$$

This two parameter model is a simple model. We can write an even simpler model in which we we set $x_i = 0$ or $x_i = 1$.

In other words,

$$\begin{aligned} y_i &= \beta_0 + \epsilon_i \\ y_j &= \beta_0 + \beta_1 + \epsilon_j \end{aligned} \tag{1.2}$$

We write this model more elegantly by writing μ for β_0 and α_i for the second term.

$$y_i = \mu + \alpha_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \tag{1.3}$$

We will impose a constraint of symmetry on the α_i's such that they are equidistant from μ as depicted in Figure 1.1.

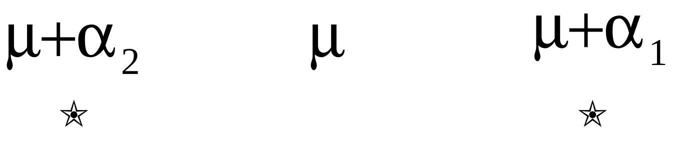


Figure 1.1

Mathematically, we impose the constraint that $\alpha_1 + \alpha_2 = 0$. This constraint will allow us to produce estimates and ANOVA tables otherwise the problem becomes indeterminable.

α_i is called a factor and it has two levels. We can specify more than two levels, like three level: α_1, α_2, and α_3 where the constraint of symmetry becomes $\alpha_1 + \alpha_2 + \alpha_3 = 0$. The number of levels can be as many as we want. Of course, the more levels, the more complicated the model. We will look at a two level factorial model.

In Figure 1.2, two factors are shown in which each factor has two levels.

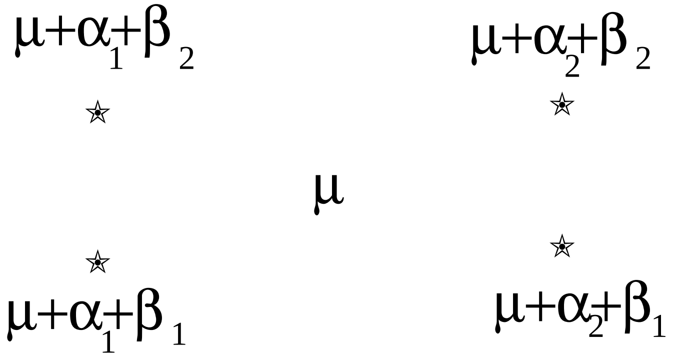


Figure 1.2

The goal is to find μ. We hope that the four corners of the rectangle bracket μ. When designing an experiment, we rely on prior experience to specify the four corners.

We will refer to the design shown in Figure 1.2 as 2×2 factorial design. In Figure 1.1, we will say that it depicts a 2 factorial design. We can generalize to more than one or two factors. A three factorial design would be written as $2 \times 2 \times 2$. A schematic diagram of it would the same as the one shown in Figure 1.2, but instead of a rectangle, the geometric figure would be something like a cube. The mathematical expression for a $2 \times 2 \times 2$ is given in equation (1.4).

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijkl} \text{ where } \epsilon_{ijkl} \sim N(0, \sigma^2) \tag{1.4}$$

As in the case of the 2 factorial design, we need to impose the constraints that $\sum \alpha_i = 0$, $\sum \beta_j = 0$, and $\sum \gamma = 0$.

The subscript i corresponds to factor α; the subscript j corresponds to factor β; the subscript k corresponds to factor γ. The subscript l denotes the replication. The experiment can be replicated several times; that is, the experiment can be done several times under the same experimental conditions to achieve greater precision in the estimates.

$$\begin{aligned}
y_{11} &= \mu + \alpha_1 + \epsilon_{11} \text{ where } \epsilon_{11} \sim N(0, \sigma^2) \\
y_{21} &= \mu + \alpha_2 + \epsilon_{21} \text{ where } \epsilon_{21} \sim N(0, \sigma^2)
\end{aligned} \tag{1.5}$$

For example, in equation (1.5), there a 2 factorial design replicated once.

$$\begin{aligned}
y_{11} &= \mu + \alpha_1 + \epsilon_{11} \text{ where } \epsilon_{11} \sim N(0, \sigma^2) \\
y_{12} &= \mu + \alpha_1 + \epsilon_{12} \text{ where } \epsilon_{12} \sim N(0, \sigma^2) \\
y_{21} &= \mu + \alpha_2 + \epsilon_{21} \text{ where } \epsilon_{21} \sim N(0, \sigma^2) \\
y_{22} &= \mu + \alpha_2 + \epsilon_{22} \text{ where } \epsilon_{22} \sim N(0, \sigma^2)
\end{aligned} \tag{1.6}$$

In equation (1.6), there is a 2 factorial design replicated twice.

$$\begin{aligned}
y_{11} &= \mu + \alpha_1 + \epsilon_{11} \text{ where } \epsilon_{11} \sim N(0, \sigma^2) \\
y_{12} &= \mu + \alpha_1 + \epsilon_{12} \text{ where } \epsilon_{12} \sim N(0, \sigma^2) \\
y_{13} &= \mu + \alpha_1 + \epsilon_{13} \text{ where } \epsilon_{13} \sim N(0, \sigma^2) \\
y_{21} &= \mu + \alpha_2 + \epsilon_{21} \text{ where } \epsilon_{21} \sim N(0, \sigma^2) \\
y_{22} &= \mu + \alpha_2 + \epsilon_{22} \text{ where } \epsilon_{21} \sim N(0, \sigma^2) \\
y_{23} &= \mu + \alpha_2 + \epsilon_{23} \text{ where } \epsilon_{23} \sim N(0, \sigma^2)
\end{aligned} \tag{1.7}$$

In equation (1.7), there is a 2 factorial design replicated three times.

By writing equation (1.7) in matrix notation, the patterns of a 2 factorial design replicated three times will be more apparent.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The design matrix has an interesting structure. The first column is always filled with 1's. The second and third columns remind us a binomial random variable in which the random variable represents two outcomes: 1-0, success-failure, on-off, up-down, pass-fail, high-low. We will write the design matrix again; this time we will ignore the first column, because we know that it is always filled with 1's. Instead of 1-0, let us write H for high and L for low.

$$\begin{bmatrix} H & L \\ H & L \\ H & L \\ L & H \\ L & H \\ L & H \end{bmatrix}$$

Actually the second column is redundant. Therefore, for the α factor, we can write the matrix shown as follows:

$$\begin{bmatrix} H \\ H \\ H \\ L \\ L \\ L \end{bmatrix} \tag{1.8}$$

By looking at this matrix, we immediately recognize a 2 factorial design with two levels and replicated three times. It suggests the nature of the experiment. We will take measurements at the high condition and at the low condition. We hope that the high level and the low level will bracket μ.

Example 1 Whenever we bake an apple pie, conditions vary especially if we do not precisely follow the recipe. In a commercial bakery, conditions needs to be highly controlled, in order to produce consistently good pies. Let us pretend that we want our home kitchen to be run like a commercial bakery. The Betty Crocker recipe book was written perhaps 60 years ago when ovens then performed differently than modern ovens. We know that the temperature of the oven will determine a good pie or a bad pie. We find a judge with good discriminatory taste to evaluate our pies on a scale of 1=bad to 5=excellent. The high temperature setting will be 400°F and the low temperature setting will be 350°F. The pies will be baked for 40 minutes. From experience, the high temperature is a little too high and the low temperature is little too low. We are confident that based on our experience, they will bracket the optimum temperature for earning a 5 from the judge.

If we were to bake one pie at 400°F and another pie at 350°F, we will have conducted a 2 factorial design experiment replicated once. Suppose on the next day, two pies were baked at the high and low temperature settings, then the experiment will have been replicated twice. Suppose on the third day, another two pies were baked at the high and low temperature settings, then the experiment will have been replicated three times. As the number of replicates increases, the more expensive the experiment becomes in terms of time, resources, and money.

From experience, we know that temperature of the oven is an important factor in determining the taste of a pie. We may design an experiment of not only setting the temperature of the oven at two different levels, but we can bake the pies at two different lengths of time; perhaps at 30 minutes and 40 minutes. Now, the experiment is a 2×2 factor design replicated once.

The experiment can be enlarged to include multiple bakers: Richard, Hank, and Sue. The experiment has become a $2 \times 2 \times 3$ factor design replicated once. There might be other possible factors. If we keep the experimental conditions as stable as possible, then we can eliminate any confounding effect. For example, though Hank and Sue might be fluent in English, Richard might only know German, so that if the instructions are written in English, Richard is likely to make mistakes in following the recipe. Therefore, to eliminate the confounding effect of the language of the recipe, Richard should be given a German translation of the recipe.

To produce reliable scores of the pies, we have to assume that the judge gives scores with the same accuracy from pie to pie.

The design of experiment must take into account not only the assumptions, but also unexpected circumstances like what should be done in the event of a power failure or if a different variety of apple is mistakenly used. ■

Following the idea of matrix (1.8), we can show the factors and levels of a 2×2 factorial design replicated once for the model: $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$ by the matrix shown here:

$$\begin{bmatrix} H & H \\ H & L \\ L & H \\ L & L \end{bmatrix} \tag{1.9}$$

When we look at the matrix shown above, we notice that there are entries for all combinations of factors and levels. It looks balances.

Suppose, on the other hand, that the matrix had looked like the one shown as follows:

$$\begin{bmatrix} H & H \\ H & L \\ L & H \end{bmatrix} \tag{1.10}$$

We see that the L-L row is missing. Not all combinations of factors and levels appear in matrix (1.10). This is an example of an unbalanced design while matrix shown (1.9) refers to a balanced design. The necessary mathematics to produce ANOVA tables for an unbalanced design becomes very sophisticated. Imagine, in the case of this unbalanced design, a corner of the rectangle show in Figure 1.2 having disappeared. The mathematics has to deal with the missing information by employing such things as generalized inverses, the nature of which conform with certain constraints of the statistician. Even though statistical software packages will produce numbers by default, the statistician must understand what mathematical constraints the software is imposing on the unbalanced design problem for producing ANOVA's. If a statistician is not careful, he may be misled by the output and make a wrong conclusion.

In Figure 1.3, the co-ordinates of the corners of the rectangle are given by y_{11}, y_{12}, y_{21}, and y_{22}. We design the experiment so as to bracket μ. Once the set of data, for example in a 2×2 factorial design, has been collected, $\widehat{\mu} = \frac{y_{11}+y_{21}+y_{12}+y_{22}}{4}$. This $\widehat{\mu}$ lies at the center of the rectangle as shown, and we hope that it is close to the true μ.

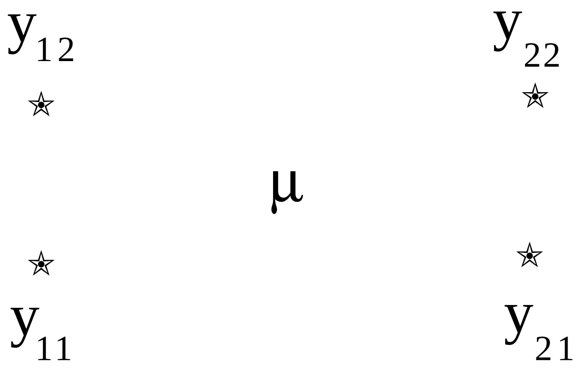


Figure 1.3

The co-ordinates of μ are $\left(\frac{y_{11}+y_{21}}{2}, \frac{y_{21}+y_{22}}{2}\right)$

Table 1.1: Signs for Calculating Effects

μ	α	β	$(\alpha\beta)$	y
+	-	-	+	y_{11}
+	+	-	-	y_{21}
+	-	+	-	y_{12}
+	+	+	+	y_{22}
Divisor 4	2	2	2	y_{22}

Table 1.1 shows a scheme for calculating estimates of the main effects and of the $(\alpha\beta)$

interaction.

$$\begin{aligned}\widehat{\mu} &= \frac{y_{11}+y_{21}+y_{12}+y_{22}}{4} && (1.11)\\ \widehat{\alpha} &= \frac{-y_{11}+y_{21}-y_{12}+y_{22}}{2} && (1.12)\\ \widehat{\beta} &= \frac{-y_{11}-y_{21}+y_{12}+y_{22}}{2} && (1.13)\\ \widehat{\alpha\beta} &= \frac{y_{11}-y_{21}-y_{12}+y_{22}}{2} && (1.14)\end{aligned}$$

$\widehat{\alpha}$ is the average of the lengths of the two horizontal edges of the rectangle. They correspond to the case of measuring the effect of α at the same level of β. For example, if α represents the effect of temperature of the oven, we measure its effect on taste both times, High-Low temperatures, at 30 minutes and then both times again at 40 minutes. α is a measure of how much temperature affects the taste of the pie at a given time. Likewise, β is the average of the lengths of the two vertical edges. β is a measure of how much time affects the taste of a pie twice, High-Low times, at a given temperature. These estimates tell us how sensitive taste is due to temperature and to time. The interaction term, $\widehat{\alpha\beta}$, tells us how much temperature and time are correlated with each other. For example, the clock on the oven might actually be affected by the heat coming from the oven in that the hotter the oven, the slower the clock as a result the $\widehat{\alpha\beta}$ interaction term will probably be a negative number and will probably be significant according to the ANOVA table.

Example 2 (Georgetown University) A set of data which was obtained from an experiment on examining the effects of levels of nitrogen and the structure of an habitat on the number of species of arthropods over a four month period on seven locations was produced by a four parameter fixed effects linear model with the response variable, y_{ijkl}, being the number of species of arthropods:

$$y_i = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl} \text{ where } \epsilon_{ijkl} \sim N(0, \sigma^2) \tag{1.15}$$

The definitions of the factors are given in Table 1.2.

A model is deemed to be a good model if the underlying theory makes sense, if there appears to be a conspicuous pattern in a plot of the data, if we can reject the hypothesis that the parameters of the model are zero, and if the assumptions of the model like the assumption that the residuals are indistinguishable from white noise are valid.

The factors which the analyst deemed to have been useful are: `Richness`, `Fert`, `Thatch`, `Month`, and `Block`.

Table 1.2: Definitions of Factors

Effect	Definition	Given Variable Name	Level	Meaning
y	Number of Species	Richness		
α	Fertilization Treatment	Fert	0 L H	None Low High
β	Habitat Structure	Thatch	0 Th	Thatch Removed Thatch Present
γ	Month	Month	1 2 3 4	17 June 27 June 12 July 12 August
δ	Block	Block	1 2 3 4 5 6 7	

`Richness` is the response variable. It is a measure of the number of species of arthropods which are found in an area of a certain size. According to the theory, if the habitat is fertile and healthy, there should be an abundance of arthropods with many species. The scientists believe that applications of fertilizer, `Fert`, will improve the growth of the flora and thereby produce a better habitat for arthropods. `Thatch` is either removed or it is left undisturbed. The effect, `Month`, takes into account changes in sunlight intensity and moisture during the growing season. Finally, there is the factor, `Block`.

Much of the vocabulary of experimental designs is derived from agricultural research which was conducted by Ronald Fisher and other British scientists. It seems obvious that growing conditions depend on soil, moisture, and sunlight, for example. However, these conditions might differ by the location of a plot of land. A plot might have a slope; another one might have less fertility; another might be sodden with water. The scientists called these plots, blocks. Generally, we are not interested whether one block is more productive than another, rather, we are interested in measuring the effect of fertilizer and the amount of residual thatch in affecting the richness of species of arthropods.

The term block is used in other experiments as a variable which has an effect on the response but which is not a subject of concern. It is used to eliminate a variable as a confounding effect. For example, does attending lecture increase Final Examination scores. A blocking variable might be the sex of the student; it might be a significant factor, but we are not interested in it. Instead, we are interested in the effect which attendance plays on examination scores regardless of a student's sex. We would block on sex as a way to eliminate it as a confounding variable.

The structure of the original set of data suggests the following four factor fixed effects linear model:

$$y_i = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl} \text{ where } \epsilon_{ijkl} \sim N(0, \sigma^2) \tag{1.16}$$

which was presented in the introduction as equation (1.15).

According to the model, we are dealing with a five dimensional set of data. As such, it is impossible to make a picture of the data at once. Instead, we will look at two dimensional slices of the data, in order to discover whether there exists any conspicuous relationships between the variables or trends in the data.

1.1 Make a Picture of the Data

To begin with, we will look at each variable individually as if there are not other factors present. The presence of other factors will muddy any patterns which might appear, but in the gross context, we might see some obvious trends.

There appears in Figure 1.4 an increase in Richness due to a high level of fertilizer over no fertilizer. We, therefore, should expect to reject the hypothesis that the richness is the same across levels of fertilizer in the ANOVA table.

Figure 1.5 suggests that removing the thatch does not affect the richness in species of the area.

Perhaps month has an affect on richness as suggested in Figure 1.6. We might see a corresponding significant effect in the ANOVA table.

A blocking factor is used to eliminate an effect. Presumably, `block` accounts for seven geographic areas. Whether there is a difference in richness between locations evidently is not a matter of concern whereas within a block the factors of fertilizer, thatching, and month are being examined to assess whether they can explain the response.

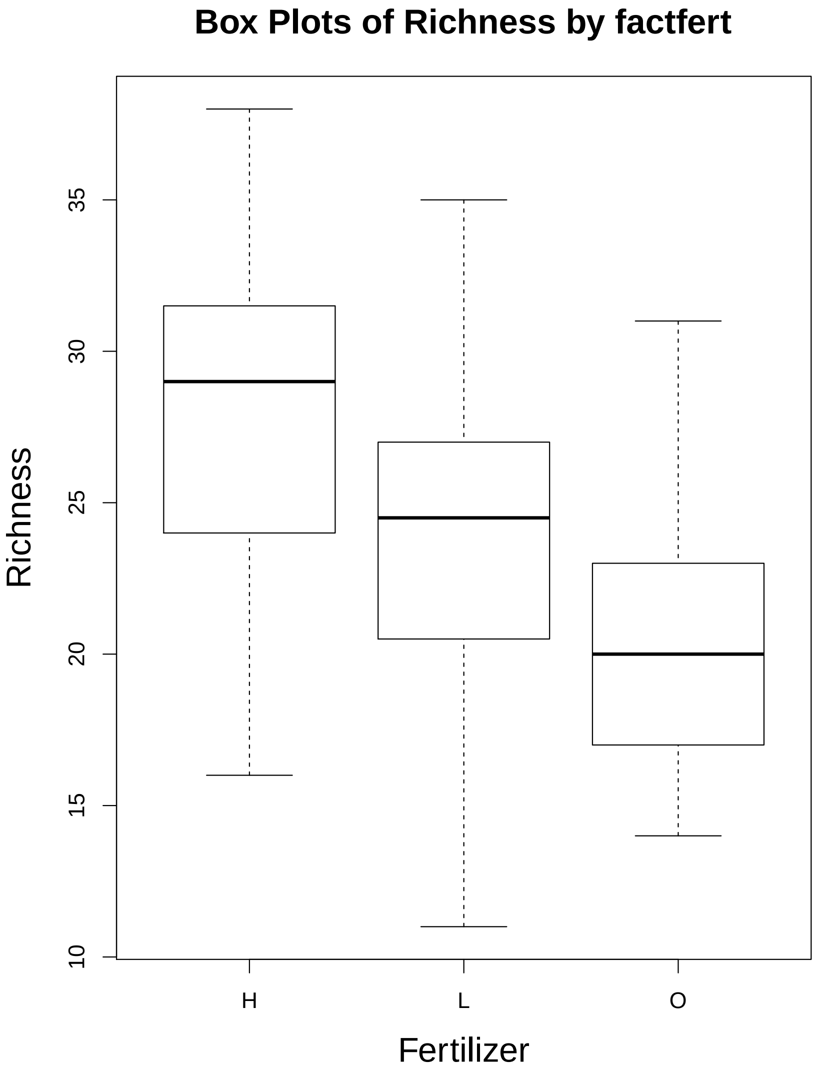


Figure 1.4: Box Plots of Richness by Levels of Fertilizer

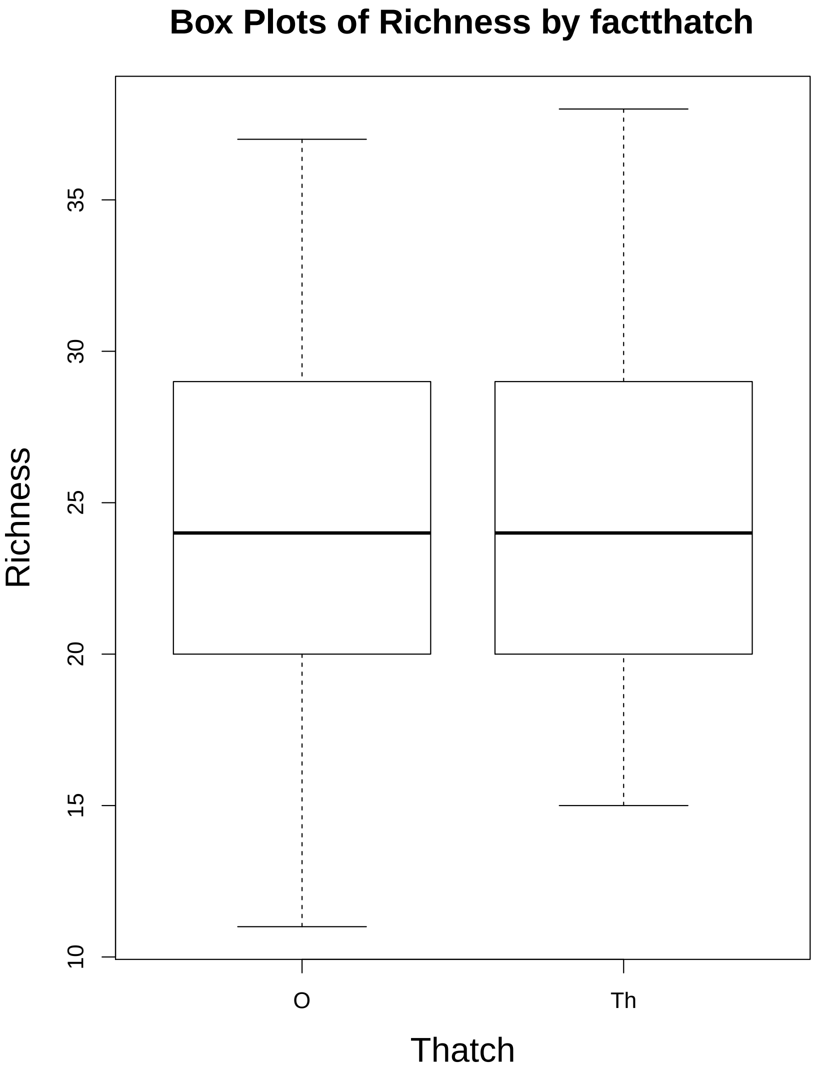


Figure 1.5: Box Plots of Richness by Levels of Thatching

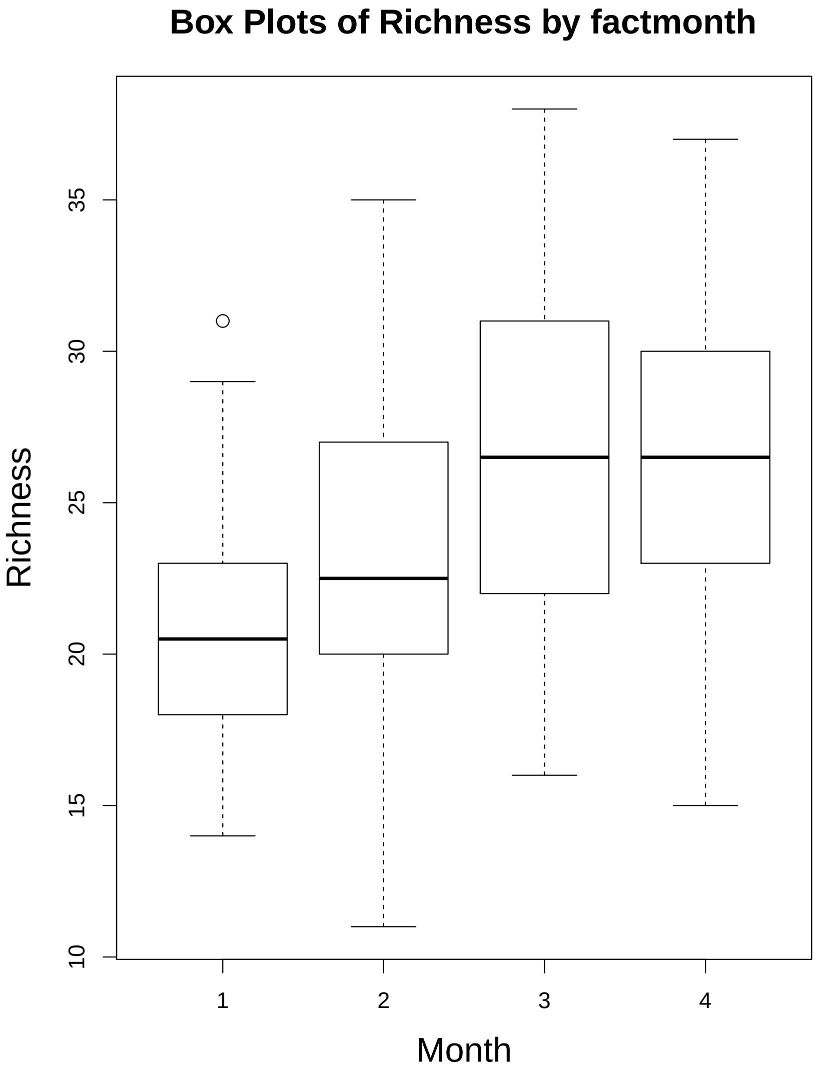


Figure 1.6: Box Plots of Richness by Levels of Month

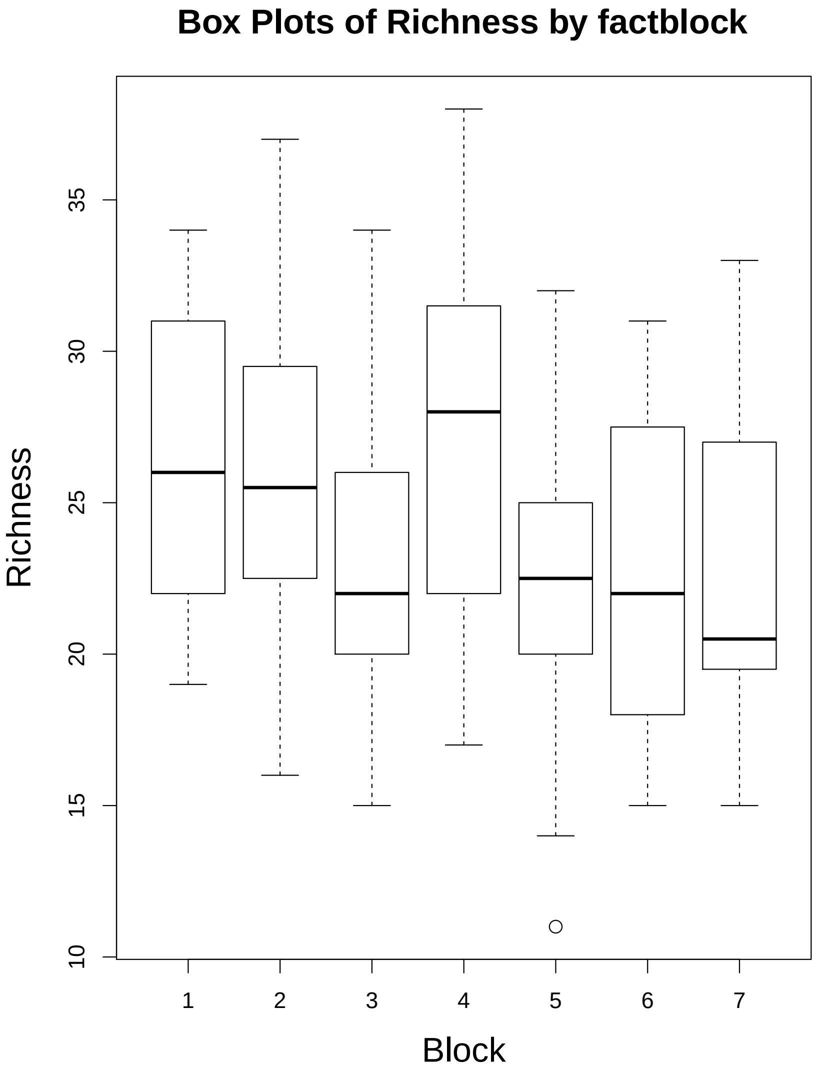


Figure 1.7: Box Plots of Richness by Levels of Blocks

1.2 Analysis of Variance Table

An analysis of variance table can be produced by means of some statistical software product. A model such as the one given by equation (1.16) and which is written again as:

$$y_i = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl} \text{ where } \epsilon_{ijkl} \sim N(0, \sigma^2) \tag{1.17}$$

forms the basis of an analysis of variance to test the hypothesis that the four factors are significant in explaining the richness in the population of arthropods. By means of the R software program, the following ANOVA is produced:

```
Analysis of Variance Table

Response: Richness
            Df  Sum Sq Mean Sq F value    Pr(>F)
    fert     2 1620.33  810.17 70.3243 < 2.2e-16 ***
    thatch   1   24.38   24.38  2.1163    0.1478
    block    6  685.24  114.21  9.9134 3.000e-09 ***
    month    3  917.79  305.93 26.5553 6.467e-14 ***
    Residuals  155 1785.67   11.52
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the factor for Thatch has a p-value of .1478. We may assume that it does not make a significant contribution in explaining the response variable, Richness.

As was expected from inspecting Figures 1.4 and 1.6, fertilizer and month are significant factors in explaining Richness. Somewhat surprising is that Block is an important factor even though according to the box plots given in Figure 1.7, there does not appear to be a conspicuous difference between them. We need to keep in mind that a plot of the data like Figure 1.7 is only a two dimensional slice of a five dimensional set of data.

In order to asses the validity of the assumption of the model that $\epsilon_{ijkl} \sim N(0, \sigma^2)$, that is, the assumption that the residuals resemble white noise. The plot of residuals versus predicted values shown in Figure 1.8 shows a random pattern; therefore, we may consider that assumption of the ϵ_{ijkl}'s is valid. Moreover, the QQ plot, also, shown in Figure 1.8 shows a good diagonal trend. Both diagnostic plots support the claim that the model is a good model.

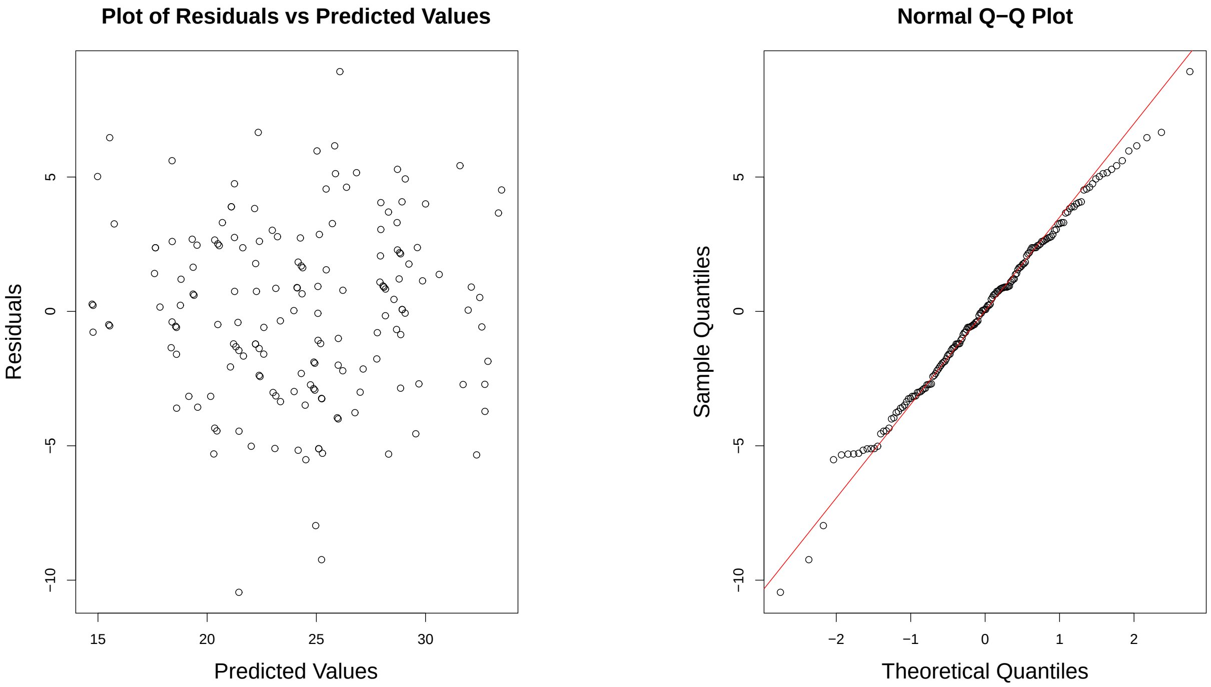


Figure 1.8: Diagnostic Plots for the Assumption of Normality of the Residuals

Confidence intervals, of course, are of paramount importance in making statistical inferences. Because the current set of data lies in a five dimensional space, five dimensional confidence regions are impossible to draw. For the same reason when a series of box plots were made to examine the data in two dimensional slices, two dimensional confidence intervals are constructed for `Richness` according to `month` and `Thatch`. Though the effects of `Block` are confounded in the confidence intervals, the confidence intervals provide a useful portrayal of the effects of `Fert`, `Thatch`, and `month` on `Richness`.

The pattern of applying fertilizer to improve richness is apparent in Figure 1.9. High fertilizer always produces a higher richness. We see in the same figure that keeping the thatch or removing it will not affect richness to which the ANOVA agrees.

In conclusion, the scientists showed that the vitality of the flora which a high level of fertilizer promotes is an important factor regardless of location, month, and the presence of thatch. ■

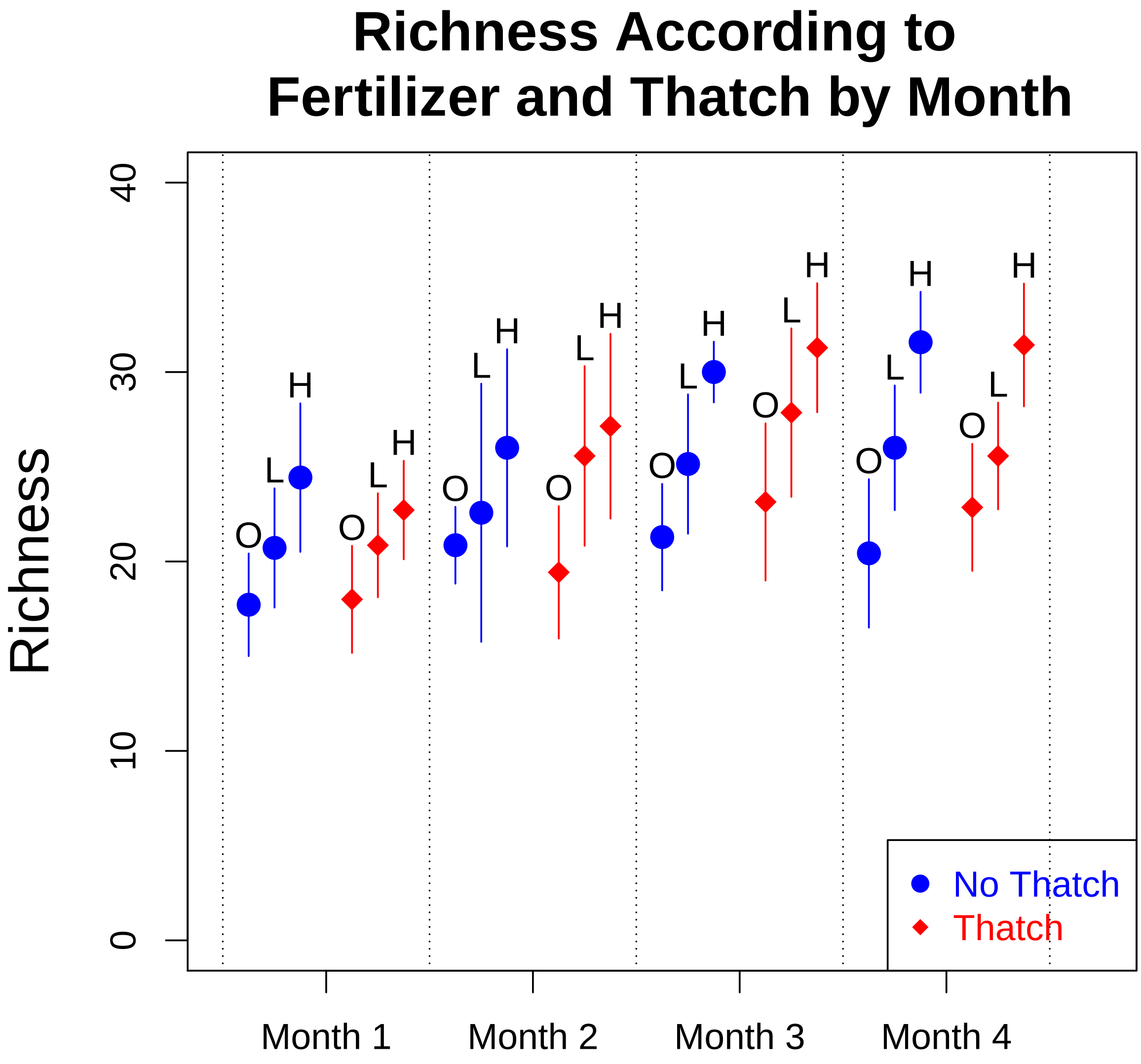


Figure 1.9: 95% Confidence Intervals of Richness by Fert, Thatch, and month.

Chapter 2

McClave Problem 9.62 Cows

Stress in cows prior to slaughter. What is the level of stress (if any) that cows undergo prior to being slaughtered? To answer this question, researchers designed an experiment involving cows bred in Normandy, France (*Applied Animal Behaviour Science*, June 2010). The heart rate (beats per minute) of a cow was measured at four different pre-slaughter phases - (1) first phase of visual contact with pen mates, (2) initial isolation from pen mates for prepping, (3) restoration of visual contact with pen mates, and (4) first contact with human prior to slaughter. Data for eight cows (simulated from information provided in the article) are given in the SPSS set of data called: `COWS.sav`.

2.1 Discussion

The model is:

$$\begin{aligned} BPM_{ij} &= \mu + cows_i + phase_j + \epsilon_{ij} \text{ where } \epsilon_{ij} \sim N(0, \sigma^2) \\ y_{ij} &= \mu + \alpha_i + \beta_j + \epsilon_{ij} \text{ where } \epsilon_{ij} \sim N(0, \sigma^2) \end{aligned}$$

for $i = 1, \ldots, 8$ and $j = 1, \ldots, 4$

The set of data was evidently obtained from an 8×4 factorial design replicated once where we are interested in determining whether or not phase makes a difference while cow is being blocked since we do not care about differences from cow to cow. The response variable is BPM which is beats per minute of a cow's heart.

To that end, we will test the hypothesis that $H_0 : phase_1 = phase_2 = phase_3 = phase_4 = 0$ vs $H_1 : otherwise$ at a level of significance $\alpha = .05$. The effect due to the cows will be removed by the analysis of variance, so that the effect of phase will not be confounded with the effect of the differences between cows. The Analysis of Variance Table is given below:

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
phase      3  521.12 173.708  3.6302 0.0296773 *
cow        7 1922.87 274.696  5.7406 0.0008201 ***
Residuals 21 1004.87  47.851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see both in Figure 2.1 and in the ANOVA table that phase has an important effect on the heart rate. Also in Figure 2.1, we see that there are differences between cows which is reflected in the ANOVA table. The plot of residuals versus predicted values as shown in Figure 2.2 seems to have a downward trend and in the QQ plot as shown in Figure 2.3 there appears to be a deviation from the diagonal at both ends of the plot. Based on these diagnostic plots of the assumption that $\epsilon \sim N(0, \sigma^2)$, the model appears to be defective.

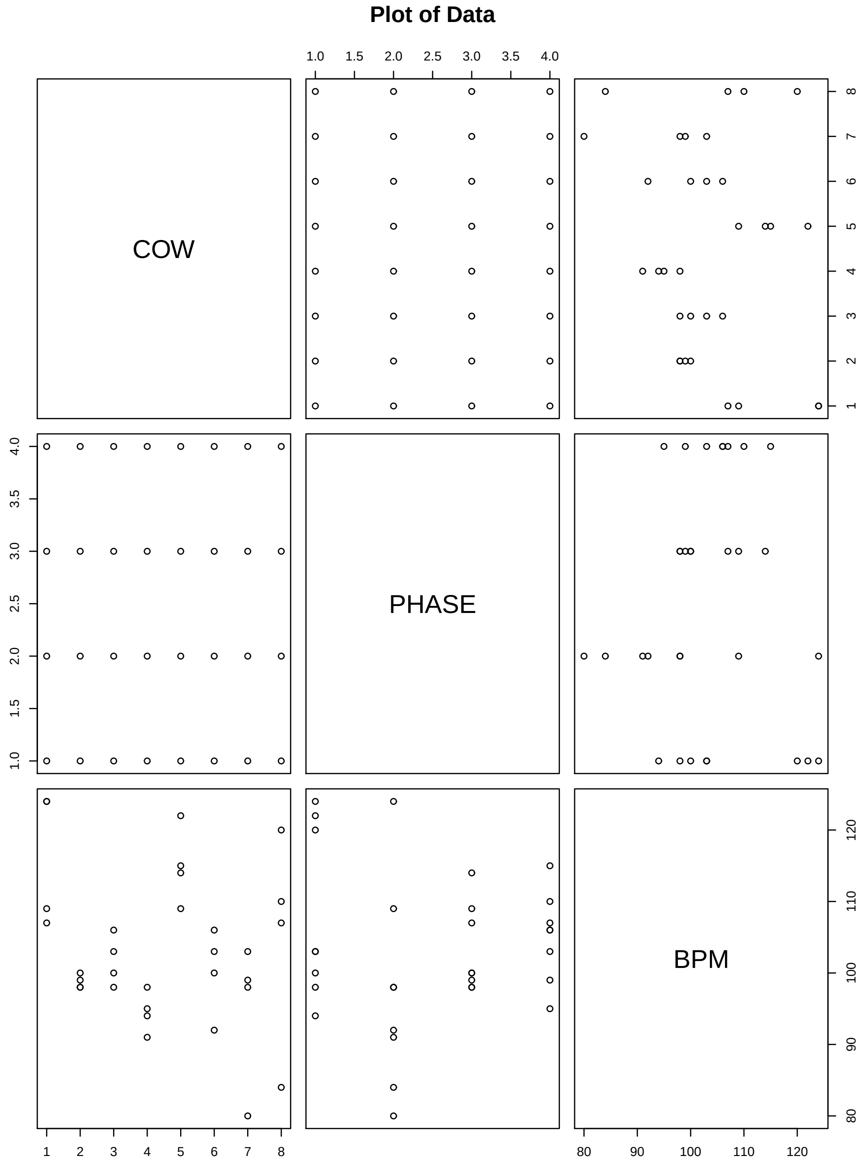

Plot of Data
COW
PHASE
BPM
1.0 1.5 2.0 2.5 3.0 3.5 4.0
1 2 3 4 5 6 7 8
80 90 100 110 120

Figure 2.1

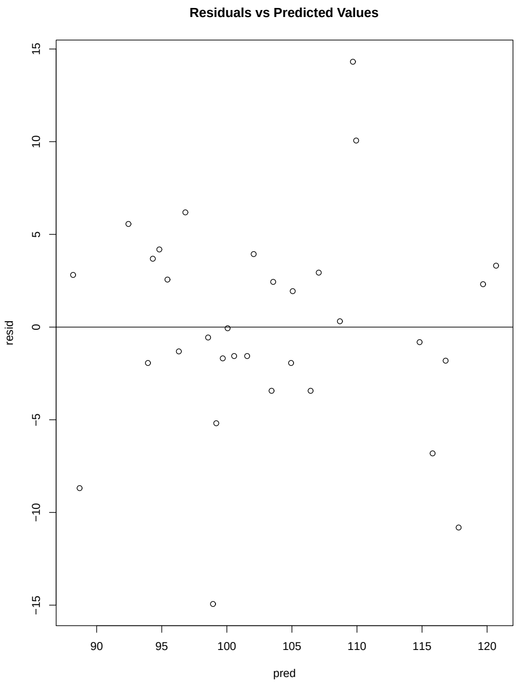


Figure 2.2

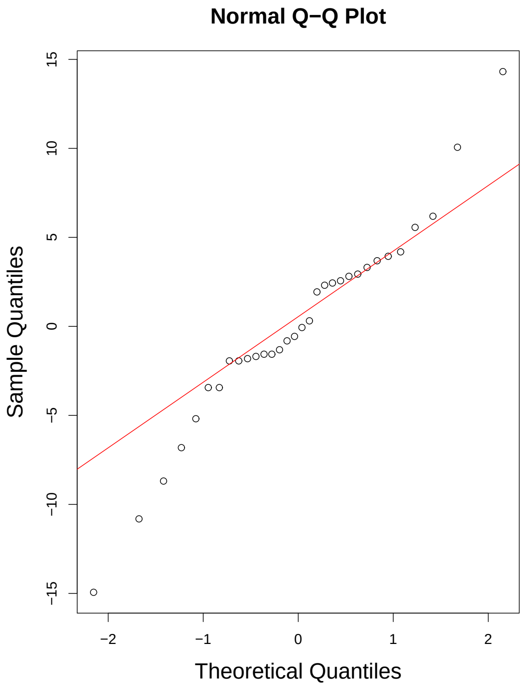


Figure 2.3

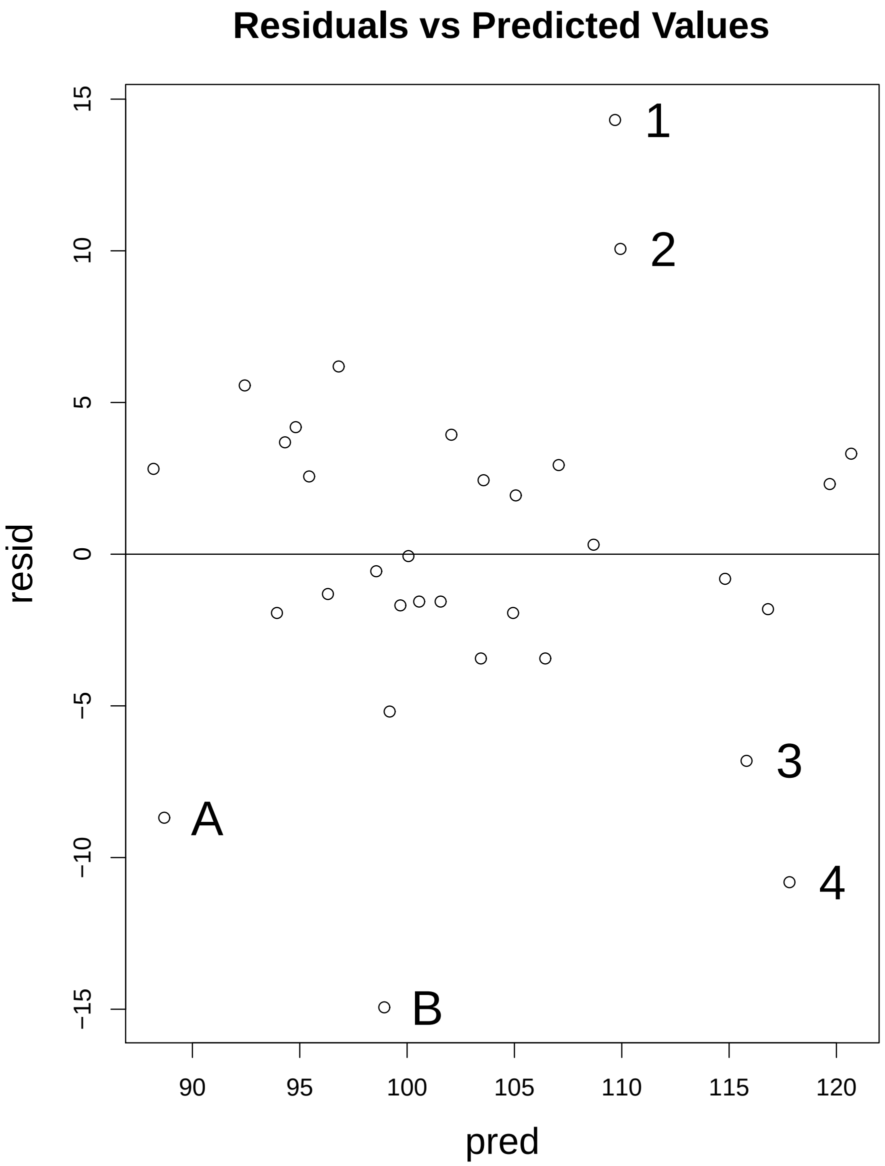


Figure 2.4

With a closer inspection of the plot of residuals versus predicted values, six points can be identified and they correspond to the points in the QQ plot which deviate from the diagonal.

It appears that cow #1 and phase #2 are not being adequately explained by the model. Either that cow is problematic and not representative of the population of cows which are suitable for slaughter or there exists a characteristic of the cow which is confounding factor which the model experimenter has not taken into account. Likewise, there is something not right with phase #2.

Point	Cow	Phase	Residual	Predicted Value
A	7	2	-8.6875	88.6875
B	8	2	-14.9375	98.9375
1	1	2	14.3125	109.6875
2	8	1	10.0625	109.9375
3	1	3	-6.8125	115.8125
4	1	4	-10.8125	117.8125

Perhaps the experimenters should examine the conditions of phase #2 to make sure that phase #2 is well defined in terms of timing or ambient conditions.

2.2 Questions

1. cow and phase are the main effects where cow is the block and phase is the effect of interest.

2. According to the ANOVA table, both effects are significance at a level of significance, $\alpha = .05$.

3. The plot of the data and the ANOVA table agree that there are differences between phases.

4. The problem with comparing confidence intervals for each of the phases is the confounding effect which cows have on them.

5. None of the questions which Professor McClave asks address the diagnostics of the model. By examining them, we learn that the experimenters need to examine more closely the behavior of cow #1 and the conditions of phase #2. Though they look like outliers, they provide useful information for the experimenters to improve their model.

Chapter 3

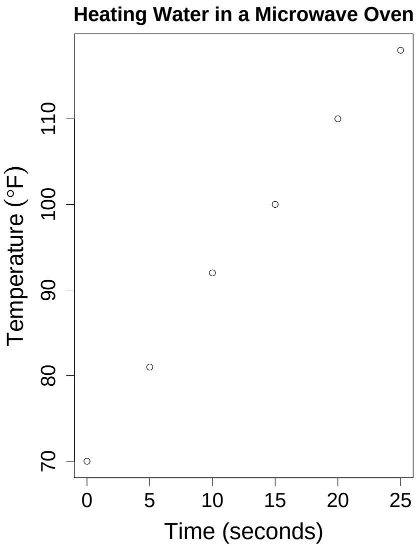
Microwave Experiment

3.1 Description of the Experiment and Proposed Linear Model

Six plastic cups are filled with 8 oz. of water each. All cups of water are allowed to come to room temperature. The first cup of water being unheated serves as the basis. The second cup is heated in a microwave oven for 5 seconds; the third cup of water is heated for 10 seconds; the fourth cup of water is heated for 15 seconds, and so on until the last cup of water is heated for 25 seconds. After each cup is heated, the final temperature of the water is recorded. The results of the experiment are shown in the table and graph below.

Table 3.1

Time (sec)	Temperature (^oF)
0	70
5	81
10	92
15	100
20	110
25	118



It should go without saying that the first order of business is to make a picture of the data. When the set of data for the microwave experiment is plotted, the trend looks linear; therefore, it seems reasonable to assert that there exists a linear relationship between temperature and the time spent in heating the water. If we let x denote time and y denote temperature, then the assertion that a linear relationship exists between x and y is expressed by: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.

3.2 Matrix Formulation of the Proposed Linear Model

Accordingly, a system of six equations in two unknowns, β_0 and β_1, for every pair (x,y) exists.

$$\begin{aligned} 70 &= 1\beta_0 + \beta_1 0 \\ 81 &= 1\beta_0 + \beta_1 5 \\ 92 &= 1\beta_0 + \beta_1 10 \\ 100 &= 1\beta_0 + \beta_1 15 \\ 110 &= 1\beta_0 + \beta_1 20 \\ 118 &= 1\beta_0 + \beta_1 25 \end{aligned}$$

Any two equations will produce a unique solution for β_0 and β_1. However, there are six equations but only two unknowns, that is, there are too many equations. We resort to the method of least squares to produce best estimates of β_0 and β_1 which will minimize the sum of squared errors, SSE.

A matrix formulation of the six equations will make the production of the estimates easier and transparent. The same information is now written in a matrix form.

$$\begin{bmatrix} 70 \\ 81 \\ 92 \\ 100 \\ 110 \\ 118 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 5 \\ 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 25 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix} \tag{3.1}$$

Recall the definition of dot project between two vectors: $(a, b)\dot{(}A, B) = aA + bB$. In matrix notation, this can be written as:

$$\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = aA + bB$$

The six equations can be written abstractly as follows

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

where $\mathbf{X}$ is call the design matrix.

By re-arranging equation (3.2), we can write:

$$\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \tag{3.3}$$

The sum of squared errors is expressed by equation(3.4)

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_6 \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix} = \sum \epsilon^2 = SSE \tag{3.4}$$

or equivalently,

$$\begin{aligned} SSE &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{3.5}$$

The x's and the y's are given by the data; they are constants. The unknowns are: β_0 and β_1. SSE is therefore a function of β_0 and β_1. To minimize $SSE(\beta_0, \beta_1)$, we differentiate SSE with respect to β_0 and β_1 and set the derivatives to zero, that is, $\frac{\partial SSE}{\partial \beta_0} = \frac{\partial SSE}{\partial \beta_1} = 0$ and solve for β_0 and β_1. In the matrix formulation of SSE as shown in equation (3.5), matrices are differentiated with respect to the vector, $\boldsymbol{\beta}$. After a series of matrix algebraic manipulations, the least squares estimator becomes:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3.6}$$

3.3 Computation Using the Matrix Formulation of the Proposed Linear Model

Equation (3.6) is essentially the starting point for deriving least squares estimates. We will apply equation (3.6) to the microwave problem. From equation (3.1)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 5 \\ 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 25 \end{bmatrix}$$

and

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 5 & 10 & 15 & 20 & 25 \end{bmatrix}$$

so that,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 5 & 10 & 15 & 20 & 25 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 5 \\ 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 25 \end{bmatrix} = \begin{bmatrix} 6 & 75 \\ 75 & 1375 \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i y_i \end{bmatrix}$$

The inverse of $\mathbf{X}'\mathbf{X}$ is:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{\begin{bmatrix} 1375 & -75 \\ -75 & 6 \end{bmatrix}}{6(437.5)} = \frac{\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}}{nSS_{xx}} \tag{3.7}$$

The final piece for the computation of equation (3.6) is:

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 5 & 10 & 15 & 20 & 25 \end{bmatrix} \begin{bmatrix} 70 \\ 81 \\ 92 \\ 100 \\ 110 \\ 118 \end{bmatrix} = \begin{bmatrix} 5716 \\ 7975 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \tag{3.8}$$

Putting the pieces together into equation (3.6), we get

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \frac{\begin{bmatrix} 1375 & -75 \\ -75 & 6 \end{bmatrix}}{6(437.5)} \begin{bmatrix} 5716 \\ 7975 \end{bmatrix} \\ &= \begin{bmatrix} 71.238 \\ 1.914 \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{y} - \widehat{\beta}_1 \bar{x} \\ \frac{SS_{xy}}{SS_{xx}} \end{bmatrix} \end{aligned}$$

The least squares fit is: $\hat{y} = 71.238 + 1.914x$ and it is drawn through set of data as shown in Figure 3.1.

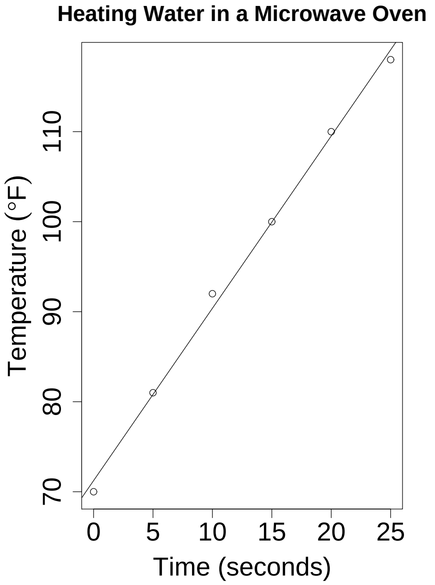


Figure 3.1

3.4 ANOVA Table and Diagnostics

In Table 3.2, the residuals are listed and their sum of squares is given as 5.6178 which is SSE. It is the same number which appears in the ANOVA table given by Table 3.3 for residual errors under the sum of squares column.

Table 3.2

Time (sec)	Temperature (^oF)	$\widehat{E[y_i]}$	$\widehat{\epsilon_i} = y_i - \widehat{y_i}$	$\widehat{\epsilon_i}^2 = (y_i - \widehat{y_i})^2$
0	70	71.238	-1.238	1.5326
5	81	80.810	$+.190$	$.0361$
10	92	90.381	$+1.619$	2.6212
15	100	99.952	$+.048$	$.0023$
20	110	109.524	$+.476$	$.2266$
25	118	119.095	-1.095	1.1990
Total			0	5.6178

Based on the F test statistic which we computed in the ANOVA table, we may reject the null

Table 3.3: Analysis of Variance Table

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	$6(95.1667)^2$=54340.17		
Regression	1	1.914(837.5)=1603.215	1603.215	F=1142
Residual Error	4	55949-54340.17-1603.215=		
		5.615	$s^2 = 1.40375$	
Total	6	$\sum_{i=1}^{6} y_i^2 = 55949$		

hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ at $\alpha = .05$, because $F = 1142 > 7.71 = F1, 4, .05$. According to the random pattern which is exhibited the first diagnostic plot shown in Figure 3.2, we may conclude that the ϵ's represent white noise in agreement with the assumption of the asserted linear model. The QQ plot shown in Figure 3.3 shows a fairly diagonal line which validates the assumption that the ϵ's follow a Normal distribution. This latter assumption justifies the use of making inferences by testing the hypothesis that $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ by means of the F test statistic and for constructing confidence intervals based on the Student's t distribution. Because both the F and Student's t distributions are derivatives of the Normal distribution, it is imperative to verify the assumption of normality of the ϵ's.

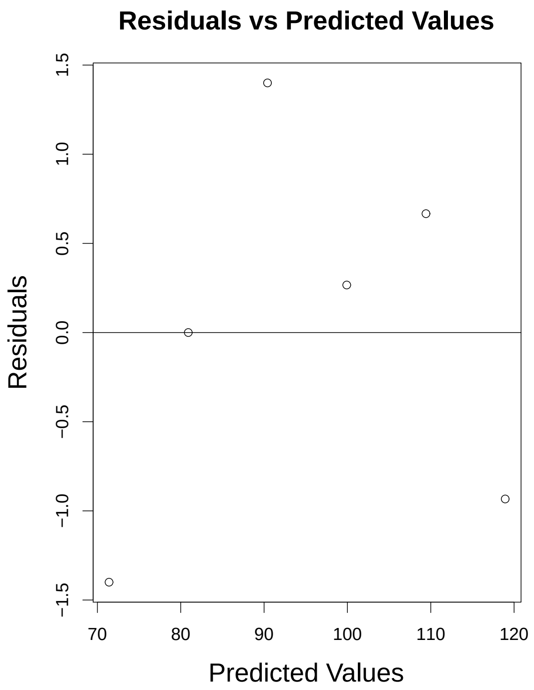


Figure 3.2

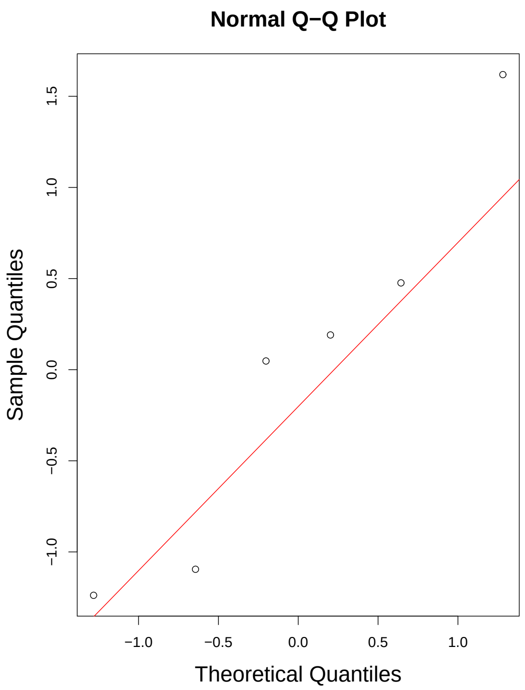


Figure 3.3

3.5 Confidence Interval for an Estimate Produced by a Linear Model

An estimate needs to have a confidence interval. To the end of constructing such a confidence interval, let $\mathbf{X}$ be a fixed $n \times p$ matrix. Each row of $\mathbf{X}$ corresponds to one observation of a vector of p explanatory variables. Let $\boldsymbol{\beta}$ be a $p \times 1$ vector of parameters for the linear model. The components of $\boldsymbol{\beta}$ are fixed, but, in general, are unknown. The linear model can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The least squares estimate of $\boldsymbol{\beta}$ is:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3.9}$$

We will use the microwave data to illustrate the construction of confidence intervals for a linear model. The design matrix and the vector of the responses for the microwave experiment are given below:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 5 \\ 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 25 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 70 \\ 81 \\ 92 \\ 100 \\ 110 \\ 118 \end{bmatrix}$$

As was done earlier, we will use equation (3.9) to produce the following estimates.

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 71.238 \\ 1.914 \end{bmatrix}$$

Given a fitted model, we use it to interpolate. The fitted model is: $\widehat{y_p} = 71.238 + 1.914x_p$ What is the confidence interval for the estimated temperature, $\widehat{y_p}$, of the water when it is heated in the microwave over for$x_p = 17$ seconds, for example?

The formulas for the lower and upper limits of the confidence intervals for $\widehat{E[y_p]}$ and for $\widehat{y_p}$ are shown in Table 3.4. Let $\mathbf{q}_p = \begin{bmatrix} 1 \\ x_p \end{bmatrix}$ so that we may write the fitted model,$\widehat{y_p} = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$ as $\widehat{y_p} = \mathbf{q}'\widehat{\boldsymbol{\beta}}$

Since $\mathbf{q}_p = \begin{bmatrix} 1 \\ x_p \end{bmatrix} = \begin{bmatrix} 1 \\ 17 \end{bmatrix}$, $\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 71.238 \\ 1.914 \end{bmatrix}$, and using equation (3.7), we get

$$\begin{aligned} \mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q} &= [1,17]\frac{\begin{bmatrix} 1375 & -75 \\ -75 & 6 \end{bmatrix}}{6(437.5)}\begin{bmatrix} 1 \\ 17 \end{bmatrix} = .21295 \\ \text{Also, } \mathbf{q}'\widehat{\boldsymbol{\beta}} &= [1,17]\begin{bmatrix} 71.238 \\ 1.914 \end{bmatrix} = 103.776 \\ t_{n-2,\frac{\alpha}{2}} &= t_{4,.025} = 2.776 \\ s &= \sqrt{1.40375} \text{ from ANOVA Table } \rightarrow s = 1.11847 \end{aligned}$$

Table 3.4: 100(1-α)% Confidence Intervals (a,b)

$\widehat{E[y_p]}$	$\widehat{y_p}$
$a = \mathbf{q}'\widehat{\boldsymbol{\beta}} - t_{n-r;\frac{\alpha}{2}} s\sqrt{\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}}$	$a = \mathbf{q}'\widehat{\boldsymbol{\beta}} - t_{n-r;\frac{\alpha}{2}} s\sqrt{1+\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}}$
$b = \mathbf{q}'\widehat{\boldsymbol{\beta}} + t_{n-r;\frac{\alpha}{2}} s\sqrt{\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}}$	$b = \mathbf{q}'\widehat{\boldsymbol{\beta}} + t_{n-r;\frac{\alpha}{2}} s\sqrt{1+\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}}$
where $s^2 = \frac{sse}{n-r}$	where $s^2 = \frac{sse}{n-r}$

By means of the appropriate equations found in Table 3.4,

$$\begin{aligned} a &= \mathbf{q}'\widehat{\boldsymbol{\beta}} - t_{n-r;\frac{\alpha}{2}} s\sqrt{\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}} \\ &= 103.776 - 2.776(1.11847)\sqrt{.21295} = 102.34 \\ b &= \mathbf{q}'\widehat{\boldsymbol{\beta}} + t_{n-r;\frac{\alpha}{2}} s\sqrt{\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}} \\ &= 103.776 + 2.776(1.11847)\sqrt{.21295} = 105.21 \end{aligned}$$

for the expected temperature, and

$$\begin{aligned} a &= \mathbf{q}'\widehat{\boldsymbol{\beta}} - t_{n-r;\frac{\alpha}{2}} s\sqrt{1+\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}} \\ &= 103.776 - 2.776(1.11847)\sqrt{1.21295} = 100.36 \\ b &= \mathbf{q}'\widehat{\boldsymbol{\beta}} + t_{n-r;\frac{\alpha}{2}} s\sqrt{1+\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}} \\ &= 103.776 + 2.776(1.11847)\sqrt{1.21295} = 107.19 \end{aligned}$$

for a particular temperature. The confidence intervals are shown in Table 3.5.

In the special case of a simple two parameter fixed effects model, the formulas for the confidence interval shown in Table 3.4 collapse to the familiar ones as shown in Table 3.6.

Table 3.5: 100(1-α)% Confidence Intervals (a,b)

$\widehat{E[y_p]}$	$\widehat{y_p}$
(102.34,105.21)	(100.36,107.19)
where $\widehat{E[y_p]} = 103.776$	$\widehat{y_p} = 103.776$

Table 3.6: 100(1-α)% Confidence Intervals (a,b)

$\widehat{E[y_p]} = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$	$\widehat{y_p} = \widehat{\beta}_0 + \widehat{\beta}_1 x_p + \epsilon_p$
$a = \widehat{E[y_p]} - t_{n-2;\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$	$a = E[\widehat{y_p}] - t_{n-2;\frac{\alpha}{2}} s\sqrt{1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$
$b = \widehat{E[y_p]} + t_{n-2;\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$	$b = E[\widehat{y_p}] + t_{n-2;\frac{\alpha}{2}} s\sqrt{1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$
where $s^2 = \frac{sse}{n-2}$	where $s^2 = \frac{sse}{n-2}$

We may plot the confidence intervals on the graph of the data as shown in Figure 3.4 in which they are shown as continuous curves. The blue pair of curves (the inner pair) correspond to the confidence intervals for the expected y, i.e. for the average while the red pair of curves (the outer pair) correspond to the confidence intervals for the particular y. Note that the curves come closest to the fitted line at the point ($\bar{x}$,$\bar{y}$). This is because when $x_p = \bar{x}$, the limits become:

$$\begin{aligned} a &= \widehat{E[y_p]} - t_{n-2;\frac{\alpha}{2}} s\sqrt{\frac{1}{n}} > \widehat{E[y_p]} - t_{n-2;\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}} \text{ for any } x_p \neq \bar{x} \\ b &= \widehat{E[y_p]} + t_{n-2;\frac{\alpha}{2}} s\sqrt{\frac{1}{n}} < \widehat{E[y_p]} + t_{n-2;\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}} \text{ for any } x_p \neq \bar{x} \end{aligned}$$

In a philosophical sense, since $(\bar{x}, \bar{y})$ lies in the midst of the data, it is surrounded by information. Usually, though, the fringes of the data contain the interesting information. For instance, a safety engineer is interested in conditions where a system will fail such as those cases which lies far from the bulk of the data or a social scientist is interested in subjects who exhibit unusual behavior like those who represent the outliers of the data. It is not uncommon to see lying with statistics by a protagonist steering the attention of the audience to the region near $(\bar{x}, \bar{y})$ and away from the fringes of the set of data.

The curves were made by using the equation

$$f(\xi) = \widehat{\beta}_0 + \widehat{\beta}_1 \xi - t_{n-2;\frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(\xi - \bar{x})^2}{SS_{xx}}}$$

for the lower limit. The curve for the upper limit follow the corresponding formula.

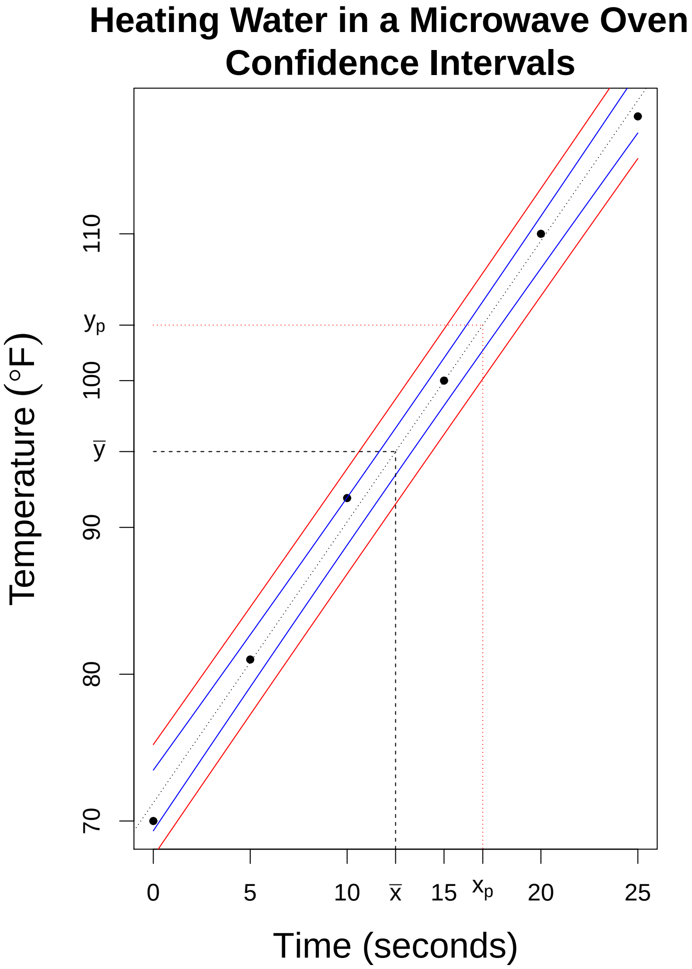


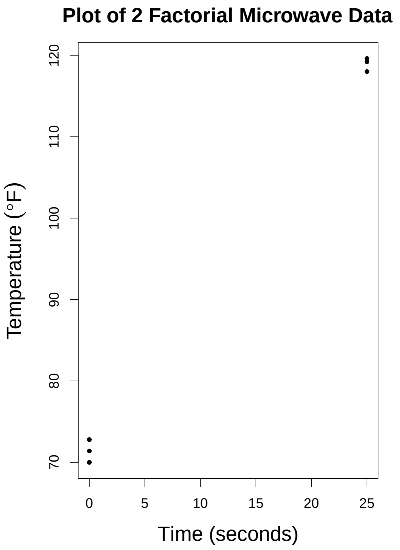
Figure 3.4

3.6 Analysis of the Microwave Data If It Had Been Gotten from a 2 Factorial Design

We noticed that temperature and time exhibit a strong linear relationship to each other. A straight line is defined mathematically by two points. A conceivable design of experiment, therefore, would be to take measurements at the two end points of the line namely at (0,70) and at (25,118). Suppose a 2 factorial design replicated three times had been followed to obtain the set of microwave data as shown in Table 3.7. In both Table 3.1 and in Table 3.7 there are measurements taken from six observations. In Table 3.1, the measurements are evenly spaced from the left to the right end points of the line while in Table 3.7 three measurements were taken at the left end point and three were taken at the right end point in accordance with a 2 factorial design replicated three times. Which design is better? Theoretically, the 2 factorial design replicated three times will produce a fitted least squares line with a smaller SSE than the design in which the observations are scattered between the end points.

Table 3.7

Time (sec)	Temperature (^oF)
0	70.0
0	71.4
0	72.8
25	119.2
25	119.6
25	118.0



The parameters of the fitted model which is based on the 2 factorial design experiment are shown in Table 3.8 along side the estimated parameters based on the original design where the observations were made at even intervals of time.

Table 3.8

Parameter	Original Design	2 Factorial Design
$\widehat{\beta}_0$	71.238	71.400
$\widehat{\beta}_1$	1.914	1.901

A tabulation of the predicted values and the corresponding residuals is shown in Table 3.9. From the table, we can make a plot of residuals versus predicted values which can be seen in Figure 3.5 along side the QQ plot shown in Figure 3.6.

Table 3.9

Time (sec)	Temperature (^oF)	$\widehat{E[y_i]}$	$\widehat{\epsilon_i} = y_i - \widehat{y_i}$	$\widehat{\epsilon_i}^2 = (y_i - \widehat{y_i})^2$
0	70.0	71.4	-1.400	1.960
0	71.4	71.4	0.0	0.0
0	72.8	71.4	$+1.400$	1.960
25	119.2	118.93	$+.266$	.071
25	119.6	118.93	$+.666$	.444
25	118.0	118.93	$-.933$	.871
Total			0	5.3066

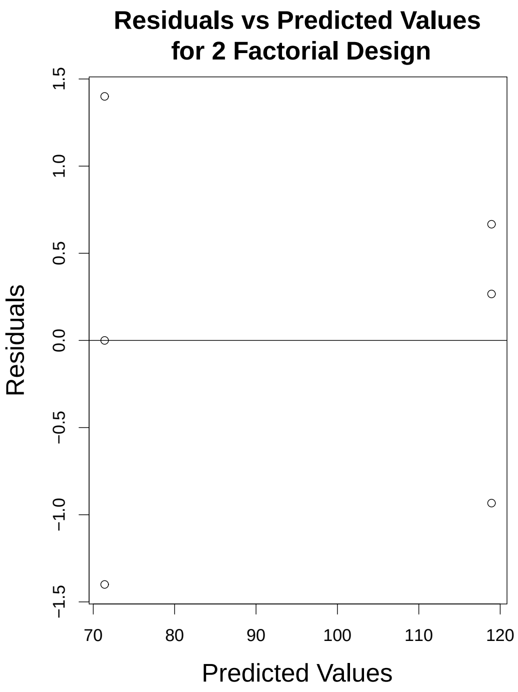


Figure 3.5

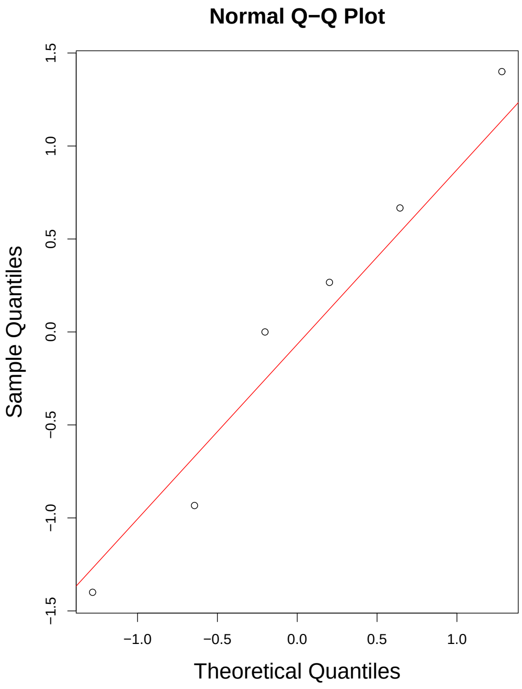


Figure 3.6

Not only is the F test statistic larger in the 2 factorial design than in the original design, but the QQ plot for the 2 factorial design looks better.

Table 3.10: Analysis of Variance Table

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	54340.17		
Regression	1	3389.127	3389.127	F=2554.618
Residual Error	4	5.3066	$s^2 = 1.3266$	
Total	6	$\sum_{i=1}^{6} y_i^2 = 57734.6$		

There are advantages for designing an experiment by collecting measurements between the end points as in the case of the original design for the microwave experiment when one does not know enough about the phenomenon to assert with confidence that the relationship between time and temperature is linear. In Figure 3.7, there appears a non-linear aberration in the functional relationship between time and temperature. It could correspond to an unexpected chemical reaction or a change in phase which would not be detected in a factorial design, but would be discovered by using the original design whereby observations are made at evenly spaced periods of time.

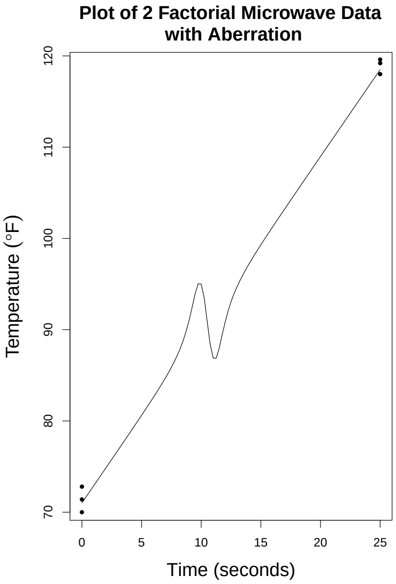


Figure 3.7