

# Chapter 6

## Adding more factors to the design

### 6.1 Introduction

In this chapter we introduce the idea of experimenting with more than one factor at a time. Many people, including many scientists I know, believe that “one at a time” experiments are the only way to really see what effect a factor has on the outcome. To introduce other factors, they say, hopelessly complicates the experiment. While, at first glance, this may seem intuitively correct, what we will show, in fact, is that not only are one at a time experiments inefficient, they are much worse. Such experiments often miss the boat completely and fail to see the most important underlying structure of the phenomenon being studied.

We’ll start with a simple example. Suppose there are two light switches on a wall and that BOTH switches have to be turned on in order to turn on the light. The engineer, who knows nothing about how the light is wired, is asked to study the effect of the two switches on the amount of light generated. Both switches start in the off position. Firmly believing that changing more than one factor at a time hopelessly confuses the results, the engineer tries the left switch first, switching it on and off many times, but, of course, finds out that this switch has no effect at all. She now moves to the right switch, and, after considerable replication, finds that this switch has no effect as well. Her report contains the conclusion that the switches are unconnected to producing light. This example may seem overly simplistic, but

let's consider a couple of real cases.

A credit card bank wants to study the effect that different interest rates and fees on a credit card have on the percentage of customers that return the application. To do this they send out two mailings: 100,000 with low interest rate and no yearly fee, and another 100,000 with a higher interest rate and a \$50 yearly fee. Not surprisingly, the response rate for the card with no yearly fee and the low interest rate is much higher – in fact, 6 percentage points higher than the response for the other card. This is a HUGE difference in this industry, one which could translate into millions of dollars of revenues. Fine. But, what they wanted me to do now was to separate out the two effects. They wanted me to tell them how much the response rate had increased due to each of the effects separately. How in the world could I do that? They had changed both of them at the same time!

Not to be fooled by changing two factors at once, another market researcher in the bank realized they weren't going to be able to separate out the effects from this design, so he decided to test for both effects, and to make things simple, he tested them one factor at a time. Since the standard card, at the time, carried a \$50/year fee, he used this in his initial mailing of 100,000 applications – 50,000 each at the two different interest rates. He found that the higher interest rate lowered the response by 1 percentage point. He then turned his attention to the fee effect using another set of 100,000 applications. Since the higher interest rate was the standard at the time, he used it, sending out 100,000 high interest rate applications, 50,000 at no fee, 50,000 at \$50/year fee. From the difference, he calculated the fee effect. It was 1.5%. From these 200,000 mailings, he was able to report both the interest rate and the fee effect, far better than his colleagues who had confounded the two. But, what did he miss? When he predicted what would happen if BOTH the fee were dropped and the interest rate lowered he predicted the response rate to increase by  $1.5\% + 1\%$  or 2.5%. But, when the other group actually tried that combination, they saw a 6% increase, more than TWICE the predicted amount! The point of this chapter will be to show how to accurately measure these effects with fewer observations. The problem here is that the effect of interest rate is not the same for the two fee levels. It's a little like the light bulb. To see the real change in response,

both “switches” had to be on!

This is one of the biggest mistakes made in industrial experimental designs today – failing to design for a possible interaction between factors. It’s one of the main motivations for my writing this book. I’ve seen it just too often. No amount of experimentation using one factor at a time will ever measure the effects of simultaneously trying different *combinations* of factors. This is the essence of an interaction effect.

But how are we to do this? We saw what happened when both factors were changes at once! What we hope to do in this chapter is to introduce designs that keep the simplicity of one at a time factor experiments, but at the same time have the ability to measure interactions. As a bonus, these designs will actually take fewer observations than running the experiment one factor at a time!

Specifically, in this chapter we will:

- Learn to design for interaction effects
- Analyze designs with two and three way interaction effects
- Learn about graphics for interactions
- Introduce the idea of transformations

## 6.2 Additivity

To start, let’s go back to gasoline additive. We’ll consider the following randomized block design, with Car type as the blocking variable, and gasoline additive as the treatment:

For the moment I’ve left one entry blank in table 6.1. What do you *expect* for the combination of Car 3 and Shell w/Lead? A plausible answer might be 35 or 36, since Car 3 seems to get about 8 mpg more than the first two under most circumstances. You might assume that this increase will stay constant across other

possible treatment levels. Suppose, however, we are told that Car 3 got 7 mpg with the leaded gasoline.

Miles Per Gallon			
Car Type	Regular Shell	Shell w/SU2000	Shell w/Lead
Car 1	21	22	26
Car 2	21	23	27
Car 3	29	31	?

Table 6.1: Data from experiment on Gasoline Additives

What happened?

At this point it might not be a bad idea to look under the car for a large pool of gasoline. Maybe the gas tank leaked. Unlikely? Perhaps the 7 was mistranscribed and should really be 37. (John Tukey has estimated that data errors of this kind occur with frequency as high as 10%!). Or, perhaps the problem is something more *persistent*, more symptomatic, like the fact that Car 3 has a catalytic converter and can not run on leaded gasoline. These are very different phenomena. The first two examples describe events that, due to some random circumstance, have caused a value that is very unlikely. The second is something inherent about the *combination* of factor levels – something that will happen every time those two factors occur at those levels. How can we tell which occurred? Sometimes, prior knowledge about the phenomenon is enough to tell us whether the effect is random or persistent. But the only way we can really distinguish them is to repeat the combination. If we repeat the experiment and get 36 mpg for Car 3 with leaded gasoline this time around, we may suspect that 7 mpg was a fluke. However, if we get 5 mpg the second time, we suspect something quite different. The persistent effect at a combination of factor levels is called an *interaction effect*. It is the effect, over and above the effects of each factor *separately* that the combination of the two factors at their specific levels have on the response. If we have  $k$  levels of one factor and  $b$  levels of another there are  $k * b$  such combinations, or separate interaction

effects. We incorporate these potential effects into our model by adding one for every combination of levels of each factor:

$$y_{tij} = \mu + \tau_t + \beta_i + \omega_{ti} + \epsilon_{tij}, \quad (6.1)$$

with, of course, the usual assumptions on the  $\epsilon_{tij}$ . Note that we now have another subscript on the observations since we must now have replication. The **interaction effect** of level  $t$  of factor A and level  $i$  of factor B is denoted by  $\omega_{ti}$ .

We can also write this model in the following form:

$$y_{tij} = \mu_{ti} + \epsilon_{tij}, \quad (6.2)$$

where

$$\mu_{ti} = \mu + \tau_t + \beta_i + \omega_{ti}. \quad (6.3)$$

This shows that the mean of each combination of factors (each **cell**) can be thought of as the sum of the effect of each factor *plus* the interaction effect. In other words, the interaction effect can be written:

$$\omega_{ti} = \mu_{ti} - (\mu + \tau_t + \beta_i). \quad (6.4)$$

This highlights the fact that the interaction is the difference between the cell mean (the mean at the combination of levels  $t$  and  $i$ ) and the sum of the other effects – the grand mean and the two main effects. To *estimate* the interaction effects, we do the obvious, substituting our estimates of each term on the right side of equation 6.4:

$$\hat{\omega}_{ti} = \hat{\mu}_{ti} - (\hat{\mu} + \hat{\tau}_t + \hat{\beta}_i) \quad (6.5)$$

$$= \bar{y}_{ti} - (\bar{y} + (\bar{y}_t - \bar{y}) + (\bar{y}_i - \bar{y})) \quad (6.6)$$

$$= \bar{y}_{ti} - (\bar{y}_t + \bar{y}_i - \bar{y}) \quad (6.7)$$

Decomposing the observations into effects now becomes:

$$y_{tij} = \bar{y} + (\bar{y}_t - \bar{y}) + (\bar{y}_i - \bar{y}) + (\bar{y}_{ti} - \bar{y}_t - \bar{y}_i + \bar{y}) + (y_{tij} - \bar{y}_{ti}) \quad (6.8)$$

$$= \hat{\mu} + \hat{\tau}_t + \hat{\beta}_i + \hat{\omega}_{ti} + \hat{\epsilon}_{tij} \quad (6.9)$$

or,

$$\begin{aligned} \text{Observations} &= \text{Grand average} + \\ &\text{Column Effects} + \text{Row Effects} + \\ &\text{Interaction Effects} + \text{Residuals} \end{aligned}$$

or,

$$\mathbf{Y} = \mathbf{A} + \mathbf{T} + \mathbf{B} + \mathbf{I} + \mathbf{R}. \quad (6.10)$$

Because the design is still balanced (since every cell has the same number of observations), the sum of squares once again add up as well:

$$S = S_A + S_T + S_B + S_I + S_R. \quad (6.11)$$

(There are ways of adjusting for unequal numbers of observations per cell, but they're beyond the scope of our discussion. We will consider only *complete* replications of the experiment in this book. But, see, for example ?? or ??.)

Let's return to our gasoline additive example of table 6.1, but now with the complete data from two replications as shown in table 6.2.

Car Type	Additive			Car Averages
	Regular Shell	Shell w/ SU2000	Shell w/ Lead	
Car 1	21,25	22,22	26,28	24
Car 2	21,23	23,25	27,31	25
Car 3	29,31	31,33	7,7	23
Treatment Avgs	25	26	21	24

Table 6.2: Data from two replications of the Gas Additive Experiment

We decompose the data into the grand average, treatment effect, block effect, interaction effect and residual (table 6.3). Note that the residual is now a *pure*

residual, or a pure estimate of the error. That is, it comes from actual replication of the experiment, not just from neglecting other terms in the model. In fact, notice that the residuals can be obtained just by subtracting the observations in each cell from their cell mean:

$$\hat{\epsilon}_{tij} = y_{tij} - \bar{y}_{ti}$$

$$\begin{pmatrix} \text{Observations} \\ 21, 25 \quad 22, 22 \quad 26, 28 \\ 21, 23 \quad 23, 25 \quad 27, 31 \\ 29, 31 \quad 31, 33 \quad 7, 7 \end{pmatrix} = \begin{pmatrix} \text{Grand Average} \\ 24, 24 \quad 24, 24 \quad 24, 24 \\ 24, 24 \quad 24, 24 \quad 24, 24 \\ 24, 24 \quad 24, 24 \quad 24, 24 \end{pmatrix} +$$

$$\begin{pmatrix} \text{Gas Additive Effects} \\ 1, 1 \quad 2, 2 \quad -3, -3 \\ 1, 1 \quad 2, 2 \quad -3, -3 \\ 1, 1 \quad 2, 2 \quad -3, -3 \end{pmatrix} + \begin{pmatrix} \text{Car Type Effects} \\ 0, 0 \quad 0, 0 \quad 0, 0 \\ 1, 1 \quad 1, 1 \quad 1, 1 \\ -1, -1 \quad -1, -1 \quad -1, -1 \end{pmatrix} +$$

$$\begin{pmatrix} \text{Interaction Effects} \\ -2, -2 \quad -4, -4 \quad 6, 6 \\ -4, -4 \quad -3, -3 \quad 7, 7 \\ 6, 6 \quad 7, 7 \quad -13, -13 \end{pmatrix} + \begin{pmatrix} \text{Residuals} \\ -2, 2 \quad 0, 0 \quad -1, 1 \\ -1, 1 \quad -1, 1 \quad -2, 2 \\ -1, 1 \quad -1, 1 \quad 0, 0 \end{pmatrix}$$

Table 6.3: Decomposition of Data into Effects

To see how this works, look at the first column in table 6.2 under “Regular Shell”. The average of the 6 observations in this column is 25. Since the grand average of all 18 observations is 24, this gives a treatment effect of  $25 - 24 = 1$ . Thus the first column of the “Gas Additives Effects” table (table 6.3) is 1. Similarly

look at the first row, corresponding to Car Type 1. The average of the 6 observations for Car Type 1 is 24, so the “effect” of Car Type 1 is estimated to be  $24 - 24 = 0$ . The other rows and columns are done similarly. Now for the interaction. Let’s look at the upper left cell, where Car Type 1 got Regular Shell. Using only the main effects (row and columns) we predict that it should get  $24 + 1 + 0$  or 25 mpg. Why? Because that’s the sum of the grand average + the column effect + the row effect. But the actual average of the mileage in this cell is 23 (from 21 and 25). So the **interaction effect** is estimate to be  $23 - 25 = -2$ . The other 8 interaction effects are calculated similarly. (You can save some time by noticing that the interaction effects sum to zero across the rows and down the columns – so you need to calculate only 4. Of course, you can save even more work by having a software package calculate it all for you!). The residuals, as usual, just clean everything up – giving back the original observations. But, as mentioned before, there is a shortcut, since the residuals are just the differences of the observations from the cell averages. Look at the Car Type 1, Regular Shell cell again. The observations are 21 and 25. Since the average is 23, the residuals are -2 and +2.

### 6.2.1 Degrees of Freedom

Now we want to compare the size of each effect to the residuals, and, as usual, we do so by comparing the mean square of each effect to the mean square of the residual. Remember that the mean squares are just the sums of squares divided by the degrees of freedom. So, we need to know how many degrees of freedom are associated with each effect. Let’s start with the grand average. Since there is only one number in the grand average table this has 1 degree of freedom as always. The gas additive has  $k = 3$  levels, and so,  $k - 1 = 2$  degrees of freedom. (One less than the number of levels, because the effects have to add to 0). Similarly, the car type (or row factor) has one degree of freedom less than its number of levels as well. Here  $b - 1 = 2$ . What about the interaction effect? There are  $9 = kb$  different interaction effects, but notice that they sum to zero both across and down the table. We really need to know only  $4 = (k - 1)(b - 1)$  of them to fill out the table. (The interaction effect, in general, has  $(k - 1)(b - 1)$  degrees of freedom.) What about

the residuals? We calculated the residuals by subtracting the observations from the means in each of the “cells”. So, in each cell, the residuals add up to zero. In other words, from each cell, I get one less degree of freedom than the number of observations in the cell. If we let  $m$  denote the number of replications, I get  $m - 1$  degrees of freedom from each cell, or since there are  $kb$  cells,  $(m - 1)kb$  degrees of freedom in total. For our example,  $m = 2, b = 3, k = 3$ , so there are 9 degrees of freedom for the residuals (one from each cell since I have only two observations in each cell).

Suppose we had only one observation in each cell – no repeats. What happens when we try to fit this model with interaction effects? (If you want to, go try this out on your favorite software). How many degrees of freedom do we get for the residual? Since  $m = 1$ ,  $(m - 1)kb = 0!$ . There are none left for residual. (Which makes an F-test pretty tricky). Where did they go? As we said in the beginning of this section, without replication it is impossible to tell the interactions from residuals. In fact, without replication, residuals ARE the interactions. With no replication, we estimate the main effects only and use the rest to estimate the error. Any interaction effects that exist are dumped where? Into the error! The residuals for the unreplicated two way design are *not* pure estimates of the error – they contain the interaction effects (if they exist) as well as random error.

### 6.2.2 The ANOVA table

To calculate the sums of squares for each effect, just go back to table 6.3, square every element in each table and take the sum. Here are the results:

What does this tell us? Assuming that our assumptions (and we should check residual plots) are correct, we would reject the hypothesis of equal gas additive means, accept the hypothesis of equal car type means and reject the hypothesis of no interaction effects.

ANOVA table					
Source of Variation	SS	df	MS	F-ratio	Significance Level
Gas Additives	84	2	42.00	14.54	.0015
Car Type	12	2	6.00	2.08	.1813
Interaction	768	4	192.00	66.44	.0000
Residual	26	9	2.89		
Total(Corrected)	890	17			

Table 6.4: ANOVA table for Additive Experiment

### 6.2.3 Interpreting effects in the presence of interactions

But, wait a minute. The ANOVA table (table 6.4), tells us that we should **accept** the null hypothesis of no Car Type effect since it has an F-value of 2.08 with a p-value of 0.1813. But, what does this mean? This doesn't seem to make sense given the data in table 6.2. Certainly which car you use affects the mileage! What's wrong? Let's look at a different way of displaying the data:

Let's again ask the question of whether the different car types affect the mileage. On the *average* Car 3 gets 23 mpg, Car 1, 24 and Car 2,

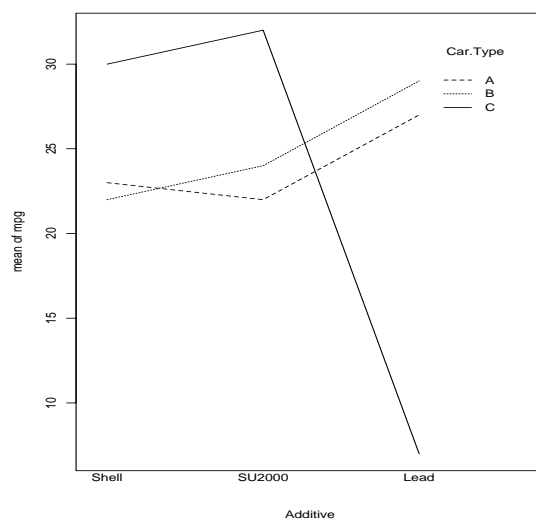


Figure 6.1: Interaction plot of mileage data. The mean value of mileage is plotted for each car type at each additive (on the x-axis). When no interaction is present, the lines are approximately parallel

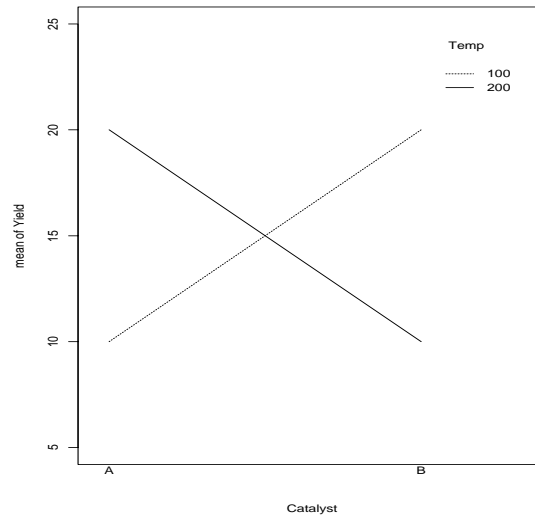


Figure 6.2: A hypothetical example showing a perfectly anti-synergistic interaction. Notice that the main effects of both temperature and catalyst are 0. The interaction effect is evident by the strong crossing of the lines

interaction effect should be tested first and the interaction plot (as in figure 6.1) examined. If the interaction is significant, all statements about main effects should be made *conditionally* on the levels of the other effect(s).

To drill this idea home, consider the experiment (with results shown in figure 6.2) to test the amount of chemical reaction of two different catalysts at two different temperatures. The response is the yield of the reaction in grams. What's the effect of temperature on the yield? Well, the main effect of temperature is estimated to be 0, since "on the average" the response at the two temperatures is the same. But to say that temperature has no effect seems to miss the boat. It reminds me of the old joke about the statistician who had his head in the oven, his feet in the freezer, but who said that he felt fine "on the average". Like with this statistician, the "average" in the case of a interaction is often nonsense. Let me reiterate: we can not interpret main effects in the presence of interactions.

### 6.3 An Example

The credit card bank of section 4.11 wants study the effect on response rate (the number of people sending back an application) of a new credit card offer *as well* as the effect of the four prices (APR interest rates). They set up the design shown in table 6.5.

Product	Annual Percentage Rate			
	Low	Medlow	Medhigh	High
Champion				
Challenger				

Table 6.5: Design for Marketing Study

Now, for each of the 8 combinations of Product and Interest Rate, they send out 10,000 offers, randomizing over the mailing list. The percentages for each cell are shown in table 6.6.

As before, we enter the number of times (the counts) for each row, as shown in table 6.7.

An ANOVA table (table 6.8) shows the interaction effect as does the interaction plot itself (figure 6.3).

Here, the interaction is obviously significant, making the main effects uninterpretable. The fact that they are both significant as well simply means that the means of, for example, the four different APR levels are different *even when averaged over*

Product	Annual Percentage Rate				Averages
	Low	Medlow	Medhigh	High	
Champion	6%	5.5%	4 %	3.8%	4.83%
Challenger	8%	7.5 %	3.5%	3.8%	5.7%
Averages	7%	6.5%	3.75%	3.8%	5.26%

Table 6.6: Data for Marketing Study

Response	Counts (F)	APR	Product
0	9400	Low	Champion
1	600	Low	Champion
0	9450	Medlow	Champion
1	550	Medlow	Champion
0	9600	Medhigh	Champion
1	400	Medhigh	Champion
0	9620	High	Champion
1	380	High	Champion
0	9200	Low	Challenger
1	800	Low	Challenger
0	9250	Medlow	Challenger
1	750	Medlow	Challenger
0	9650	Medhigh	Challenger
1	350	Medhigh	Challenger
0	9620	High	Challenger
1	380	High	Challenger

Table 6.7: Data entry for Marketing Study in *JMP*®

ANOVA table					
Source of Variation	SS	df	MS	F-ratio	Significance Level
Product	1.531	1	1.531	30.88	2.751e-08
APR	17.954	3	5.985	120.694	0.0000
Product * APR	2.594	3	0.865	17.437	2.594e-011
Residuals	3966.370	79992	0.0496		
Total(Corrected)	3988.449	79999			

Table 6.8: ANOVA table for Market Study

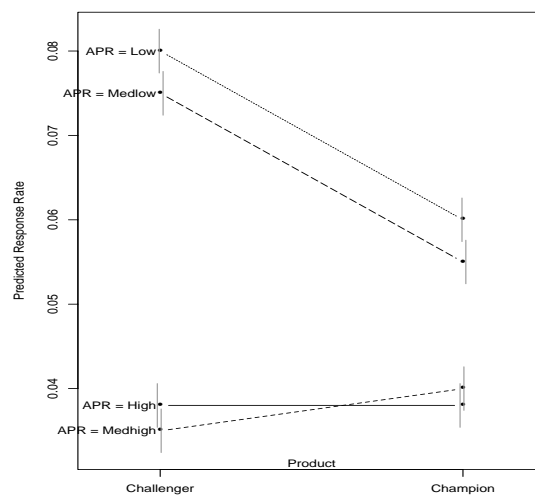


Figure 6.3: Interaction plot for the response rate, with 90% confidence limits for each combination of apr and Product. Note that for the two high levels of Apr, the two products perform quite similarly, while for the two lower levels of Apr, the challenger significantly outperforms the champion. Main effects plots for the average balance transfer. It appears that cardtype and apr have large effects.

the two products. Similarly the response rates of the two products are different even when averages over the four different APRs. But, it's the interaction that's important. Notice that for the two high levels of APR, the two products perform quite similarly (with perhaps an edge to the Champion), but that for the two low levels of APR, the Challenger significantly outperforms the Champion. Thus the decision to go with the Challenger will have to be made in context of which APR level is used. (Other costs associated with the Challenger and with switching over to it from the Champion should of course be factored into this decision as well). One of the advantages of 0-1 data is that we don't have to consider residual analysis in the same way. If the sample sizes are reasonably large, the assumption of normality for the percentages is automatically satisfied.

Where do we go from here? We could go adding additional factors. The only restriction on this strategy is the number of combinations that this might generate. A design in three factors, with 2 levels for the first factor, 3 levels for the second and 4 levels for the third would require  $2 * 3 * 4 = 24$  runs even unreplicated. The analysis of such designs is absolutely straightforward from what we have already learned. The addition of a third factor introduces the notion of a three way interaction, but the concept is similar. A three way interaction can be viewed as the effect at a combination of three factors over and above the three main effects, and the three two way interactions at those levels. In other words, it's the difference in how the two way interaction behaves at different levels of the third factor. An example will help and will be presented below. But, clearly, these designs are of limited applicability because of the number of combinations involved. In Chapter 7, we turn our attention to designs in more than 2 factors, but with only two levels. In fact, we will see that we can study as many as seven different factors in only 8 experimental runs. These designs, known as fractional factorial designs, are widely used in industry and are invaluable during the initial stages of an investigation. After narrowing down the field to two or three factors, one can study these in more detail in further experiments.

creative	apr	cardtype	avgBT
No Miles	Low	Silver	753
Miles	Low	Silver	933
No Miles	High	Silver	1360
Miles	High	Silver	1392
No Miles	Low	Gold	1603
Miles	Low	Gold	1666
No Miles	High	Gold	1865
Miles	High	Gold	1837
No Miles	Low	Silver	978
Miles	Low	Silver	1131
No Miles	High	Silver	1499
Miles	High	Silver	1555
No Miles	Low	Gold	1787
Miles	Low	Gold	1849
No Miles	High	Gold	1908
Miles	High	Gold	1979

Table 6.9: Data from experiment in three factors

## 6.4 An example with three factors

The data in table 6.9 shows a very different experiment by the credit card bank. Here they are interested in the effect of several factors on the amount of money that is transferred to the new credit card from the cardholders' other credit card accounts. This is known as a balance transfer. The factors to be studied are two credit card types (cardtype), two prices (apr), and whether the card is linked with an airlines Miles program(creative). The experiment is replicated over two different mailing lists, which are treated as two replicates of the design. The response is the average balance transfer of the cardholders who receive each type of card.

Let's look at some plots first. The main effects are shown in figure 6.4. It

appears that card type is quite important and apr as well. However, we should look at the interactions shown in figure 6.5. Notice that there appears to be a cardtype:apr interaction. The amount of balance transfer increases less for the gold cards than the silver when apr is lowered. The other plots seem roughly parallel, indicating little evidence for interactions. We now turn to the ANOVA table for confirmation of what we've seen (table 6.10).

Indeed, the two main effects of apr and cardtype are significant as well as their interaction. Because the interaction is significant but relatively weak, the main effects of apr and cardtype are still significant as well. But because of the interaction effect, we should be careful in talking about the "apr" or the "cardtype" effect by itself. The amount that apr increases or lowers the average balance transfer depends on the cardtype.

ANOVA table					
Source	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
creative	1	21683	21683	1.541	0.250
apr	1	453939	453939	32.270	0.0005
cardtype	1	1496341	1496341	106.370	0.0000
creative:apr	1	6683	6683	0.475	0.510
creative:cardtype	1	4001	4001	0.284	0.608
apr:cardtype	1	110058	110058	7.824	0.023
creative:apr:cardtype	1	1661	1661	0.118	0.740
Residuals	8	112539	14067		

Table 6.10: ANOVA table for 3 factor experiment

How can we summarize the findings of the experiment? Raising the apr charged lowers the average balance transfer, but less so for the Gold card than for the Silver. Having airline miles attached to the card does not seem to influence the amount of balance transferred. How much does the balance transfer change for different combinations of apr and cardtype? The plot 6.5 shows 90% confidence intervals

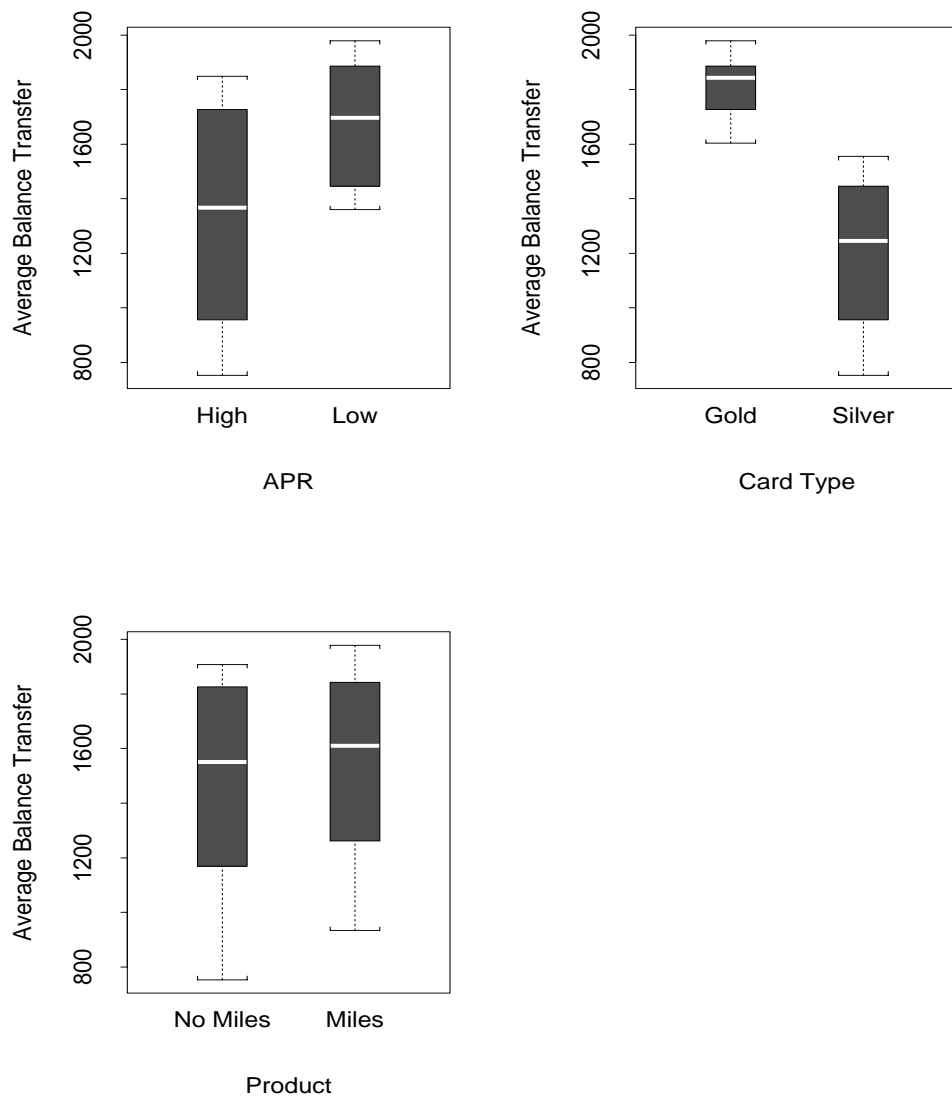


Figure 6.4: Main effects plots for the average balance transfer. It appears that cardtype and apr have large effects.

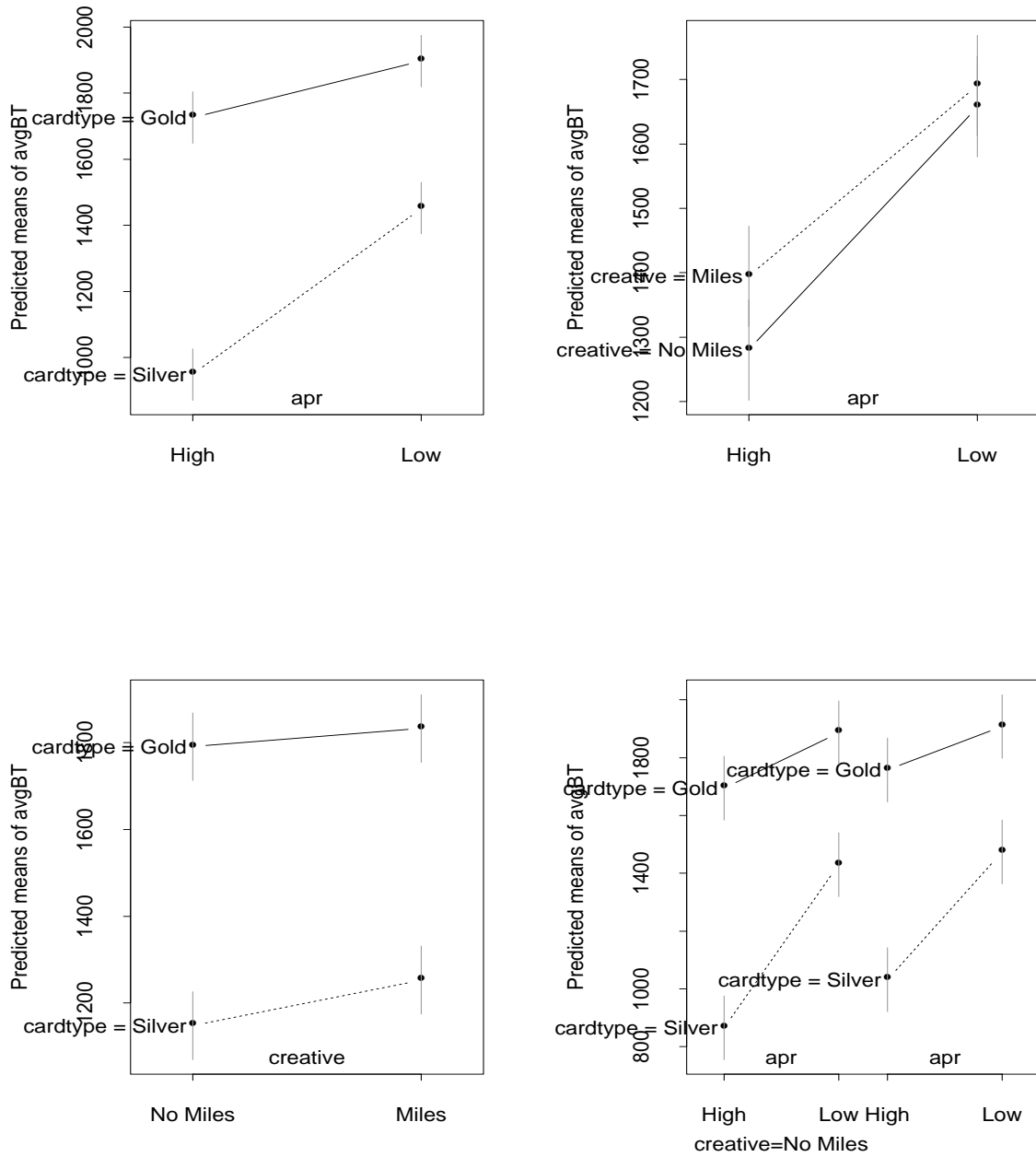


Figure 6.5: Interaction plots for the average balance transfer. It appears that there may be an apr:cardtype interaction.

for each of the four combinations of apr and cardtype (the plot in the upper left). A table of the four combinations is shown in table 6.11.

	Silver Card	Gold Card
Low Apr	(1321.09,1581.91)	(1766.84,2027.66)
High Apr	(818.34,1079.16)	(1595.84,1856.66)

Table 6.11: Confidence intervals for the 4 combinations of apr and cardtype. The lower and upper 95% limits for the average balance transfer are shown for each combination.

Of course, residual analysis should be performed as well, in order to justify the F-tests and the confidence interval analysis. Figure 6.6 shows that the residuals exhibit no particular deviations from the assumptions. The normal probability plot is a bit S shaped, however. This indicates that the residuals are a bit two short tailed (*i.e.* they don't have enough extreme values on either end) to be perfectly normal. But, the F-test is relatively robust against such deviations, especially when the residuals are symmetric like these.

## 6.5 Transformations

Let's try another example, this time a simple two factor design, replicated 10 times. Two types of gasoline, SU-2000 and regular were tested on three cars: a Lamborghini Corsach, a Chrysler Voyager and a Geo Metro. A summary of the data are shown in table 6.12, where for each combination of additive and car model, the mean and standard deviation of the mileage are given. The complete data are contained on the diskette or at the web site:

`\\http:www.williams.edu\Mathematics\faculty\deveaux\book\data.`

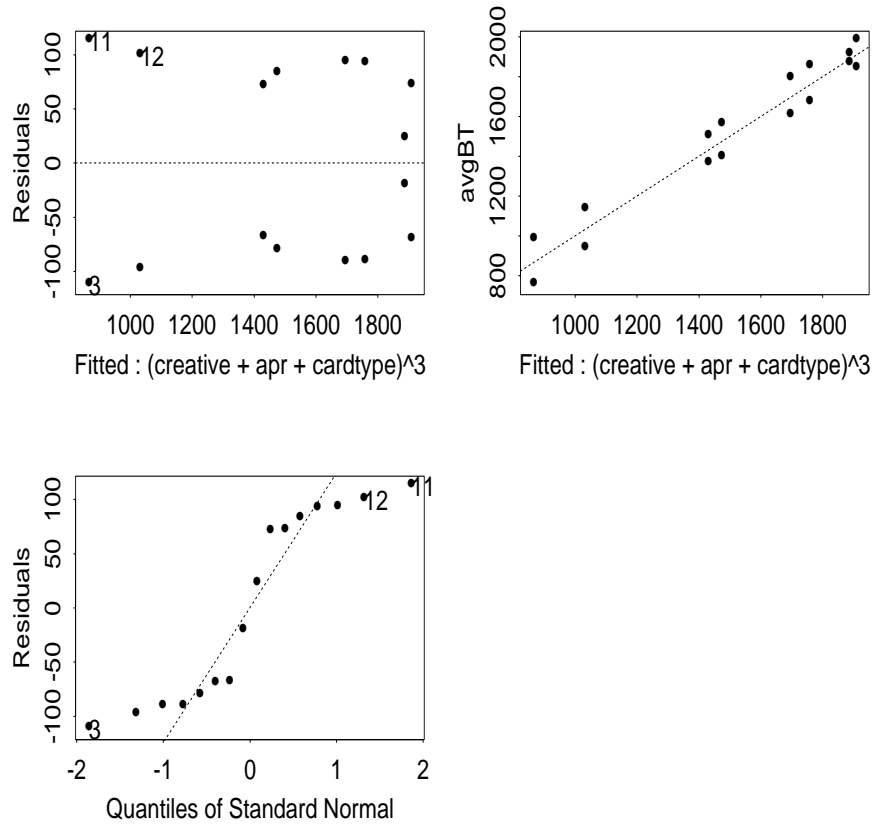


Figure 6.6: Various residual plots for the analysis of balance transfer. Note that the residuals exhibit no particular deviations from the assumptions, with the possible exception of the normal probability plot which shows that the residuals have tails that are a bit short.

Model	Additive	Mean Mileage from 10 runs	Standard Deviation of the 10 runs
Lamborghini	Regular	5.003	0.341
Lamborghini	SU-2000	5.539	0.335
Voyager	Regular	20.221	1.350
Voyager	SU-2000	22.125	1.748
Metro	Regular	50.081	3.589
Metro	SU-2000	55.067	2.773

Table 6.12: Summary of data from experiment on additives. Data listed are the summaries of the 10 runs.

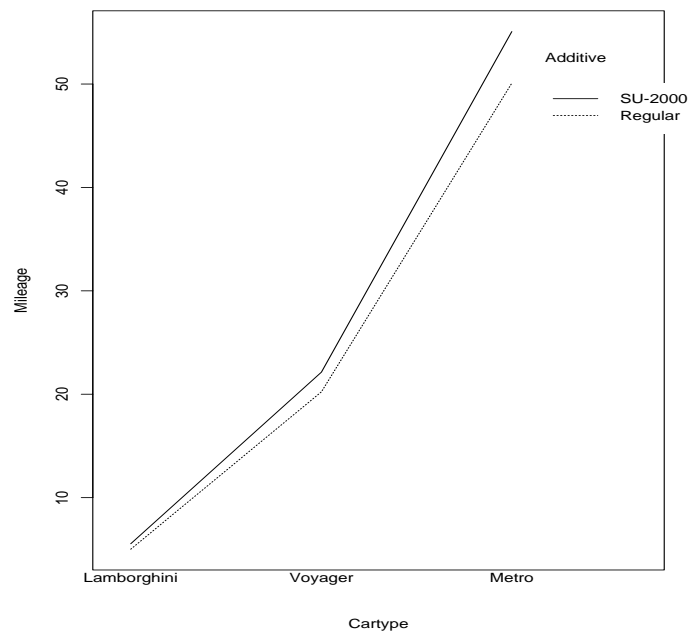


Figure 6.7: Interaction plot of mileage data.

The ANOVA table (table 6.13) looks pretty straightforward. The interaction is clearly significant as are the main effects by themselves.

ANOVA table					
Df	Sum of Sq	Mean Sq	F Value	Pr(F)	
model	2	23175.44	11587.72	2712.211	0.0000
additive	1	92.16	92.16	21.570	0.00002
model:additive	2	51.90	25.95	6.074	0.00418
Residuals	54	230.71	4.27		

Table 6.13: ANOVA table for additive experiment of table 6.12

An interaction plot shows us why (figure 6.7). The higher the mileage of the car, the more the additive helps. The amount the additive increases mileage *depends* on the model – it is not constant and hence the interaction effect. But wait, there's trouble here. Let's look at the residual plot (figure ??). The error certainly doesn't look constant. In fact, the error seems to increase proportionally to the mileage of the car. The plot thickens! Why is this bad? Let's look at the standard errors of the cell means. Since the mean square error in table 6.13 is 4.27, the estimate of the error standard deviation is  $\sqrt{4.27} = 2.07$ . Since each cell has 10 observations, the standard error of a cell mean is  $s/\sqrt{n} = 2.07/\sqrt{10} = .654$ . This is shown in table 6.14.

Let's use this table to construct a confidence interval for the mean mileage of a Lamborghini using Regular gas. Since the standard error is 0.654 we'll use the average  $\pm 2 * 0.654$  for a rough 95% confidence interval. This becomes (3.695, 6.311). For a Lamborghini using SU-2000 it's (4.230, 6.847). So, we can't statistically distinguish the effect of the additive for the Lamborghini. The intervals overlap too much. Doesn't this seem peculiar? Can't we be more precise than (3.695, 6.311) for the *mean* of a Lamborghini with regular gas? In fact, for the 10 runs the range of the Lamborghini using regular was only 4.41 to 5.34. The standard deviation of these 10 runs is only 0.34 – about half of the estimated standard error of its mean!! What's going on? Let's look at the other extreme – the Metro with SU-2000. Here

Cell	Mean	Std. Error
Lamborghini, Regular	5.003	0.654
Lamborghini, SU-2000	5.539	0.654
Voyager, Regular	20.221	0.654
Voyager, SU-2000	22.135	0.654
Metro, Regular	50.081	0.654
Metro, SU-2000	55.067	0.654

Table 6.14: Summary of the data by cell again, but this time the standard error of the cell mean is shown. Notice that all the standard errors are the same, by virtue of the same sample size and the fact that there is one common estimate of the error standard deviation. Compare this to table 6.12.

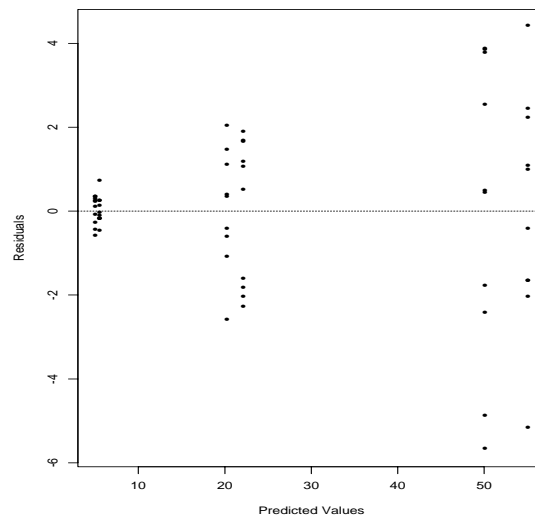


Figure 6.8: Plot of residuals versus predicted values. Notice the heteroscedasticity of the residuals. The plot thickens!

we got a range of 49.89 to 59.47 –nearly 10 mpg. The standard deviation of these 10 runs is 2.77 – over 8 times bigger than the 10 Lamborghini runs. But in table 6.14, **both** cell means have the same standard error – 0.654. Why? Because the ANOVA estimates only **one** overall error standard deviation. It can't take into account the fact that cars that get 50 mpg vary more than cars that get 5 mpg. This is the problem. It makes all confidence intervals, and in fact the F-tests themselves invalid. Moreover, what's this interaction effect anyway? Isn't there a simpler explanation here than to say that the effect of the additive depends on the mileage of the car? Look at the effect – for the Lamborghini, it increases mileage about .5 mpg, for the Voyager about 2 mpg and for the Metro about 5 mpg. Isn't there a simpler way to describe this? Yes! They all get about a 10% increase.

Statisticians often transform the response variable in a situation like this. The transformation may help to stabilize the variance, express the response in a simpler way, make the errors more normally distributed, or some combination of the three. A full discussion of how to find the appropriate transformation is beyond the scope of this book. The interested reader is referred to [?], chapter 7 for a discussion of the Box-Cox family of transformations. Many software packages now automatically check for transformations to improve the analysis. Try taking the log of the mileage for the data we have just analyzed to see the possible results of such a transformation.