2

Analysis of Full Factorial Experiments

This chapter details how to analyze 2^k factorial experiments and is organized as follows:

2.1 Analysis Strategy Overview

The following four-step strategy is recommended for the analysis of 2^k factorial experiments.

2.1.1 Step 1: Study the variation in y

Begin with a histogram of the response data y and observe the range and distribution of values. If the distribution is evenly spread, then fitted models will not be overly affected by just a small subset of the data. If the distribution

R.W. Mee, *A Comprehensive Guide to Factorial Two-Level Experimentation*,
DOI: 10.1007/b105081_2,

is severely skewed, or there are a few values far removed from the others, then the fitted models will attempt to account for this prominent variation while largely ignoring the rest.

The shape of the distribution of y can be altered by the use of a nonlinear transformation. Section 2.7 explains how such transformations may be employed to find a satisfactory simpler model, to stabilize the error variance, or to emphasize the variation at the lower or upper end of the range for y. If the treatment combinations are replicated, then one should examine the within-treatment-combination variation to check for consistency.

In addition to plotting the data, one should understand how the actual y values were obtained, since this may provide insight regarding the error variation. How large is the measurement error variance for the measurement system involved? Does the variability in y increase or decrease as the mean for y increases? Is y a count, a ratio, a standard deviation, or some other statistic? Section 2.8 provides guidance for each of these cases.

2.1.2 Step 2: Fit a "full" model

Step 2 begins by fitting a "full" model. For most situations, this will be the full factorial model (1.4). Rather than fitting a simpler model from the start and assuming it to be adequate, we prefer to fit a complex model and so confirm what terms are not needed. There are exceptions [e.g., for cases of missing treatment combinations (Section 2.10) or with prior knowledge that certain interactions are not needed] where it is preferred to begin with a simpler model. However, the typical initial model for analyzing 2^k experiments will be the full factorial model (1.4).

How we proceed after fitting a complex model will depend on whether the experiment includes replication—that is, were runs repeated at some or all of the treatment combinations? Sections 2.2–2.5 will discuss methods and tools appropriate for the different cases that arise. The objective is to determine which terms are useful for explaining the variation in y and providing insight into the factor effects.

2.1.3 Step 3: Find an adequate simpler model

Now fit a reduced (i.e., simpler) model, as appears reasonable following Step 2. The purpose here is not to determine the significance of the remaining terms but rather to perform diagnostics to determine whether the reduced model adequately explains the variation in the response (see Section 2.6). If the residual analysis indicates problems, then some remedy is required, such as adding terms to the model, questioning aberrant y_i values, or transforming the response. Once a satisfactory model is obtained, one may proceed to Step 4.

2.1.4 Step 4: Interpret and utilize the final model(s)

Use graphs to summarize the results of the satisfactory model. Express the conclusions in the natural units for each factor and the response. If a transformation for y was involved in the analysis, quantitative results should also be expressed in terms of the original measurement rather than simply on the transformed scale. If two competing models seem reasonable, compare them to see in what respects they differ. For instance, do they differ regarding the preferred level for each factor? Do their predicted values differ at the treatment combination(s) of interest?

2.2 Analysis of Numerical Responses with Replication

As in Section 1.3, here we consider the simplest (although not necessarily common) case, where the 2^k treatment combinations of a full factorial are each replicated n times in a manner that yields $N = n2^k$ observations with independently distributed errors. Section 1.3 discussed t-tests for individual coefficients, as well as a test involving all the saturated model's coefficients.

Following tests for individual coefficients, one proceeds in Step 3 of the analysis strategy to fitting a reduced model with, say, r coefficients, including the intercept, b_0, with $1 < r < 2^k$. Let $\mathbf{X}_{\text{red}}$ denote the $N \times r$ model matrix, let $\mathbf{b}_{\text{red}}$ denote the vector of least squares estimates for the reduced model

$$\mathbf{b}_{\text{red}} = (\mathbf{X}_{\text{red}}{}'\mathbf{X}_{\text{red}})^{-1}\mathbf{X}_{\text{red}}{}'\mathbf{Y} = \mathbf{X}_{\text{red}}{}'\mathbf{Y}/N,$$

and let $\hat{\mathbf{Y}} = (\hat{y}_1, \ldots, \hat{y}_N)'$ denote the vector of predicted values

$$\hat{\mathbf{Y}} = \mathbf{X}_{\text{red}}\mathbf{b}_{\text{red}}.$$

The partitioning of the sum of squares corresponding to this reduced model is given in Table 2.1.

Table 2.1. Analysis of variance for a reduced model

Source	df	SS
Model (reduced)	$r-1$	$\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$
Lack-of-fit	$2^k - r$	$\sum_{i=1}^{N}(\bar{y}_i - \hat{y}_i)^2$
Pure error	$N - 2^k$	$\sum_{i=1}^{N}(y_i - \bar{y}_i)^2$
Total (corrected)	$N-1$	$\sum_{i=1}^{N}(y_i - \bar{y})^2$

Table 2.1 expands Table 1.5, in that the saturated model's degrees of freedom and sum of squares are partitioned into two parts: the reduced model and lack-of-fit. The reduced model captures variation explained by the reduced model. Lack-of-fit contains variation that is explained by the saturated model but which is missed by the reduced model. In settings such as this, most statistical software will construct two F-tests:

- **Lack-of-fit test.** This is a test that the reduced model is adequate (i.e., that it explains all the systematic variation in the y_i values). The test statistic is
$$F_{\text{lof}} = \text{MS}_{\text{lof}}/\text{MS}_{\text{pe}},$$
where MS_{lof} and MS_{pe} denote the mean squares for lack-of-fit and pure error, respectively, computed from Table 2.1. The degrees of freedom for this test are $\nu_1 = 2^k - r$ and $\nu_2 = N - 2^k$, and the p-value is $P(F_{\nu_1,\nu_2} > F_{\text{lof}})$. A small p-value indicates that at least one of the β's for terms omitted from the model is not zero; in this case, one should consider adding terms. A large p-value indicates that the reduced model is consistent with the observed data.
- **Reduced model test.** This is a test of significance for the terms in the reduced model. It is computed as
$$F_{\text{red}} = \text{MS}_{\text{red}}/\text{MSE},$$
where the denominator is the mean square error (MSE) for the reduced model obtained by pooling lack-of-fit and pure error as follows:
$$\text{MSE} = \frac{\text{SS}_{\text{lof}} + \text{SS}_{\text{pe}}}{N - r}.$$
This MSE combines MS_{pe}, an estimate for σ^2 based on replication, with MS_{lof}, an estimate for σ^2 that is dependent on the assumption that the reduced model is correct. A small p-value is an indication that the model is useful for explaining variation in the y_i's, or, equivalently, that at least some of the β's corresponding to terms in the model are not zero.

If both F_{lof} and F_{red} have large p-values (e.g., more than 5 or 10%), then the factors have no recognizable effect on $E(y)$.

We illustrate these F-tests for the first Huhtamaki experiment, taking the additive model (1.2) as our reduced model. The resulting lack-of-fit test is

Source	df	SS	MS	F_{lof}	p-value
Lack-of-fit	4	4289.64	1072.41	3.96	0.0463
Pure error	8	2165.48	270.69		
Total error	12	6455.12			

This test, which is statistically significant at $\alpha = .05$ indicates that this simple model does not account for all the systematic variation in dry crush resistance. Hence, one or more of the four omitted interactions is active. The corresponding F-test for the significance of the fitted reduced model is

Source	df	SS	MS	F_{red}	p-value
Model	3	8,526.43	2,842.14	5.28	0.0149
Error	12	6,455.12	537.93		
Total (corrected)	15	14,981.55			

Note that because the MS_{lof} is nearly four times the MS_{pe}, the MSE is inflated by the systematic variation in MS_{lof}, reducing the size of F_{red} as well as any t statistics computed as

$$t = b_s/(\text{MSE}/N)^{1/2}. \tag{2.1}$$

If there are sufficient degrees of freedom from replication, then it is safer to just use (1.14) rather than (2.1). Here, with both F-tests statistically significant, we would conclude that the additive model is useful but that it can be improved by the addition of interaction terms.

In summary, replication of the factorial treatment combinations serves two purposes. First, it provides information about the error variance. Replication at each treatment combination yields MS_{pe} as an estimate for σ^2 and provides some ability to check the assumption that $\text{Var}(\epsilon)$ is constant across the experimental region (something we will explore later). In addition, replication at the 2^k treatment combinations increases the precision for each estimated coefficient. When the error variance is substantial, experiments with small N may have too little power to detect effects of practical importance. The issue of sample size to achieve sufficient power is relevant for every application, and is addressed in Section 13.1.

2.3 The Inclusion of Centerpoint Replicates

Taking $n > 1$ replicates at every treatment combination, as was assumed in Section 2.2, can be quite costly, especially if there are four or more factors. One option to economize on runs is to collect replicates at only a subset of the treatment combinations (Dykstra 1959). However, such unequal replication forfeits the orthogonality of the columns of $\mathbf{X}$ and so complicates the analysis. Section 2.9 will address how to analyze unbalanced factorial designs in general. Now consider an alternative economical approach to replication.

2.3.1 Centerpoint replication with all factors quantitative, with Example 2.1

When all of the factors are quantitative, an alternative to replicating some or all of the 2^k treatment combinations is to perform replicate runs at the center of the design. Replication at the center does not improve the precision of estimates for factorial effects, but it serves two other purposes. First, collecting data at the center provides a check on linearity of the factor effects. If the model is to be used for interpolation, this check is critical. If the centerpoint

runs indicate severe nonlinearity, then one often augments the design with additional treatment combinations to support estimation of a full second-order model. See Chapter 12 for details.

As with any true replication, centerpoint replication provides an estimate for σ^2. Runs at the center do not affect the orthogonality of a design and so do not cause the complication that arises from partial replication of factorial treatment combinations. This method is recommended for estimating σ^2, provided: (i) all the factors are quantitative, (ii) the constant variance assumption for ϵ is reasonable, and (iii) an unreplicated 2^k provides enough precision for estimating factorial effects.

Example 2.1: 2^5 with seven centerpoint runs

Consider now a five-factor example from Bouler et al. (1996). The experiment was conducted to improve the compressive strength of a calcium phosphate ceramic intended as a bone substitute. Biphasic calcium phosphate (BCP) is a blend of two materials denoted HA and β-TCP. BCP containing pores with diameter ≥ 100 μm promotes bone formation but generally has reduced strength. The purpose of the experiment is to create stronger BCP with such macropores. The factors and their levels are presented in Table 2.2.

Table 2.2. Factors and levels for Bouler et al.'s (1996) ceramic experiment

		Levels		
Factors		-1	0	1
x_1	HA in BCP (%)	45	60	75
x_2	Weight of naphthalene (%)	30	45	60
x_3	Diameter of macropores (μm)	100	300	500
x_4	Isostatic compaction (kPa)	1090	1630	2180
x_5	Sintering temperature (^{o}C)	900	1000	1100

The $2^5 = 32$ factorial treatment combinations were performed without replication; that is, $n = 1$. In addition, $n_0 = 7$ samples were made at the coded treatment combination (0, 0, 0, 0, 0). Bouler et al.'s (1996) work does not mention any randomization of order in preparing or testing the samples. The observed compressive strengths ranged from 0 to 59.1 mPa. Table 2.3 presents the results for all $2^5 + n_0 = 39$ runs, sorted by compressive strength. Note that 10 of the 39 samples showed no compressive strength, including all 8 combinations with $x_2 = 1$ and $x_3 = -1$; that is, all combinations with a high weight of the smallest-diameter naphthalene. Clearly, this combination is not satisfactory.

Table 2.3. Bouler et al.'s (1996) ceramic strength data

Sorted Run	x_1	x_2	x_3	x_4	x_5	Strength (mPa)
1	−1	1	−1	−1	−1	0.0
2	1	1	−1	−1	−1	0.0
3	1	1	1	−1	−1	0.0
4	−1	1	−1	1	−1	0.0
5	1	1	−1	1	−1	0.0
6	−1	1	−1	−1	1	0.0
7	1	1	−1	−1	1	0.0
8	1	1	1	−1	1	0.0
9	−1	1	−1	1	1	0.0
10	1	1	−1	1	1	0.0
11	−1	1	1	−1	−1	2.0
12	−1	−1	−1	−1	−1	2.2
13	−1	−1	−1	1	−1	2.9
14	1	1	1	1	−1	3.3
15	−1	1	1	1	−1	4.2
16	1	−1	−1	−1	−1	5.1
17	−1	−1	1	−1	−1	6.5
18	1	−1	−1	−1	1	7.0
19	1	−1	1	−1	1	7.0
20	1	−1	1	−1	−1	8.0
21	0	0	0	0	0	10.8
22	0	0	0	0	0	11.5
23	−1	−1	1	1	−1	11.7
24	0	0	0	0	0	11.8
25	1	−1	1	1	−1	12.3
26	1	−1	−1	1	−1	12.9
27	0	0	0	0	0	13.0
28	−1	1	1	1	1	13.2
29	0	0	0	0	0	13.4
30	0	0	0	0	0	13.9
31	1	1	1	1	1	14.1
32	0	0	0	0	0	14.5
33	1	−1	1	1	1	16.7
34	−1	1	1	−1	1	17.8
35	−1	−1	−1	1	1	23.4
36	1	−1	−1	1	1	25.7
37	−1	−1	1	−1	1	46.0
38	−1	−1	−1	−1	1	48.3
39	−1	−1	1	1	1	59.1

Table 2.4. Analysis of variance for Bouler et al. (1996) data

Source	df	SS	MS
Model (full factorial)	31	7034.15	226.908
Lack-of-fit (nonlinearity)	1	18.22	18.224
Pure error	6	11.12	1.853
Total (corrected)	38	7063.49	

As in Table 2.1, we construct an ANOVA with partitions for model, lack-of-fit, and pure error (see Table 2.4). Since we have fit a model containing all (linear) main effects and interactions, the lack-of-fit term has just 1 df, and is a check for nonlinearity (or curvature) of the factor effects. The nonlinearity sum of squares is based on the difference between the average response at the n_0 centerpoint replicates and the average at the N factorial treatment combinations:

$$\mathrm{SS}_{\mathrm{nonlin}} = \frac{(\mathrm{Mean@Center} - \mathrm{Mean@Factorials})^2}{n_0^{-1} + N^{-1}}. \tag{2.2}$$

Here, the mean strength for the 7 centerpoint replicates and 32 factorial treatment combinations are 12.7 and 10.92, respectively, and (2.2) equals 18.22. This lack-of-fit is small compared to the variation explained by the model ($\mathrm{MS}_{\mathrm{model}} = 226.9$), but is large compared to pure error ($\mathrm{MS}_{\mathrm{pe}} = 1.85$). Thus, while this lack-of-fit test is statistically significant ($F_{\mathrm{lof}} = 9.83$; $p = .02$), accounting for this minimal curvature would make little difference in our predicted values in the interior of the design region. Since the centerpoint mean exceeds the average at the factorial points, a model that ignores this curvature will give slightly pessimistic predictions near the center of the experimental region.

In addition to F-tests, software will report *R-square* (a.k.a. the *coefficient of determination*) from the ANOVA for a fitted model:

$$R^2 = \mathrm{SS}_{\mathrm{Model}}/\mathrm{SS}_{\mathrm{Total}}.$$

R^2 is the proportion of total variation that is explained by the fitted model. Here, the full factorial model's $R^2 = 7034.15/7063.49 = 0.996$, which is very high, reflecting the practical insignificance of the lack-of-fit.

Results similar to Table 2.4 are typical for processes with little error variation. If the pure error mean square is close to zero, virtually every term of a saturated model will be statistically significant. Here we might question whether the variation at the centerpoint replicates accurately reflects the true run-to-run error variation. The error term ϵ consists of errors from several sources, including the following:

- Measurement error in the testing of compressive strength

- Inhomogeneity of the materials used
- Other uncontrollable errors at each stage in the process of mixing, compressing, and heating the ceramic

For true replication, all of these sources affect each observation independently. However, if a shortcut was taken by preparing as a batch the material for all seven centerpoint specimens, then these specimens may vary less in strength than would be the case if this step were performed seven times, once for each specimen. Whether this is the case or not, the nonlinearity in compressive strength observed is not large enough to make a substantial difference to the fitted model.

Both the design and analysis for Bouler et al.'s (1996) experiment warrant further discussion. From reading their work, it appears that the materials may have been prepared in larger batches and then sampled to apply the compaction and temperature levels. If this is the case, the distribution of the ϵ_i's is affected, which alters how one should analyze the data (see Sections 3.5 and 3.6). Further, the fact that one-fourth of the observations showed no measurable strength calls into question using a single linear model for strength based on all the data. If zero strength indicates that the ceramic powder did not bond, then perhaps the 10 observations with $y_i = 0$ should be handled differently when constructing a model for strength. We return to this example later (Section 4.3) to address these issues.

2.3.2 Centerpoint replication with one or two qualitative factors

How can we replicate economically when some of the factors are qualitative? If all factors but one are quantitative, then collect centerpoint runs for the quantitative factors at both levels of the qualitative factor. For instance, Ellekjaer, Ilseng, and Naes (1996) conducted a cheese processing experiment in which just one of the factors, melting salt, was qualitative. They included 6 center runs—3 with melting salt A and 3 with melting salt B—along with the 32 factorial runs. If there are only two qualitative factors, one might collect one or two centerpoint runs at each of the four qualitative factor level combinations.

2.4 Analysis of Numerical Responses Without Replication

2.4.1 Model-dependent estimators for σ^2, with Example 2.2

Many two-level full factorial and fractional factorial experiments are run without any replication. In such cases, one can still produce useful estimates for the error variance σ^2, but these estimates are model dependent; that is, some assumptions must be made about the model in order to estimate the error variance. Three general approaches have been used, which depend on slightly different assumptions:

1. **Mean Square Error From an Assumed Model**: Prior to data analysis, assume a model less than the saturated model. Use the MSE from this fitted model to estimate σ^2. Provided all omitted terms have true coefficients of zero, this yields an unbiased estimator for σ^2.
2. **Mean Square Error From Final Fitted Model**: Analyze the data, and arrive at a satisfactory reduced model. Use the MSE for this model to estimate σ^2. Here, the MSE is an unbiased estimator for σ^2 only if the nonzero β's are so large that one always selects the same model. On subsequent pages, RMSE = $\text{MSE}^{1/2}$ is the acronym for *root mean square error* and will serve as an estimator for σ.
3. **Robust Estimator for σ From Saturated Model**: Fit a saturated model and use the estimates nearest to zero to construct an estimate for σ. We will use Lenth's (1989) estimator (explained below). Here one assumes that a majority of the terms for the saturated model have true coefficients that are zero. This assumption is known as *effect sparsity*.

Approach 1 for estimating σ^2 is valid, provided the assumed model is correct. For example, with an unreplicated 2^4 factorial, we might assume that no three-factor or higher-order interactions exist and fit the model (1.3). The resulting ANOVA will have 10 df for the model and 5 df for error. Provided $\beta_{1\cdot2\cdot3} = \beta_{1\cdot2\cdot4} = \beta_{1\cdot3\cdot4} = \beta_{2\cdot3\cdot4} = \beta_{1\cdot2\cdot3\cdot4} = 0$, the MSE is a valid estimator for σ^2.

Although approach 2 is commonly used in practice to estimate σ^2, it is the most subjective method and entails dangers that can make it unreliable. For instance, if one naively selects a model by excluding only a few of the smallest estimates (e.g., using backward elimination regression), the MSE for the reduced model will generally be much smaller than σ^2. As a result, many inactive terms may appear statistically significant.

We now introduce Lenth's method and compare it with the first two methods for estimating σ^2, using data from a chemistry experiment.

Example 2.2: Unreplicated 2^4 isatin experiment

Consider now the data from Davies (1954, p. 275). This 2^4 experiment involved a laboratory investigation of yield for a derivative of the chemical isatin. Table 2.5 lists the four factors of interest to the chemist and the levels used in this initial investigation. The 2^4 treatment combinations were each performed once in random order. Table 2.6 lists the yield in grams per 10 g. of base material. The range 6.04 – 6.79 for yield seems rather small. Because this is an initial investigation into the process, the chemist had no knowledge of σ^2 but believed that three-factor and higher-order interactions were unlikely.

Table 2.5. Factors and levels for isatin yield experiment (Davies 1954)

		Levels	
Factors		−1	1
x_1	Acid strength (%)	87	93
x_2	Reaction time (min)	15	30
x_3	Amount of acid (mL)	35	45
x_4	Reaction temperature (°C)	60	70

Table 2.6. Coded treatment combinations sorted by isatin yield

x_1	x_2	x_3	x_4	Yield
1	−1	−1	−1	6.04
1	1	−1	1	6.08
−1	−1	−1	−1	6.08
1	−1	1	−1	6.09
−1	1	1	−1	6.12
1	1	1	1	6.23
−1	−1	1	−1	6.31
1	1	1	−1	6.36
1	−1	1	1	6.38
1	1	−1	−1	6.43
−1	1	1	1	6.49
−1	1	−1	−1	6.53
1	−1	−1	1	6.68
−1	1	−1	1	6.73
−1	−1	1	1	6.77
−1	−1	−1	1	6.79

These data are used to illustrate the potential advantages and disadvantages of the three methods for estimating σ^2:

1. Assuming away higher-order interactions, we fit model (1.3) with four main effects and six two-factor interactions, and obtain the analysis of variance

Source	df	SS	MS	F
Model	10	0.8525	0.08525	2.216
Error	5	0.1923	0.03847	
Total (corrected)	15	1.0448		

Diagnostics (as explained in Section 2.6) for this fitted model show no outliers or systematic patterns. Although the usefulness of this model is questionable, given $F = 2.216$ (p-value = .20), the MSE = 0.038 provides a valid estimate for σ^2, provided the true regression coefficients for the five higher-order interactions are zero. The resulting t-tests are as follows:

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	6.3819	0.0490	130.16	<.0001
x_1	−0.0956	0.0490	−1.95	.1086
x_2	−0.0106	0.0490	−0.22	.8370
x_3	−0.0381	0.0490	−0.78	.4720
x_4	0.1369	0.0490	2.79	**.0384**
x_1x_2	−0.0006	0.0490	−0.01	.9903
x_1x_3	0.0169	0.0490	0.34	.7447
x_1x_4	−0.0806	0.0490	−1.64	.1610
x_2x_3	−0.0331	0.0490	−0.68	.5293
x_2x_4	−0.1256	0.0490	−2.56	.0505
x_3x_4	−0.0131	0.0490	−0.27	.7996

With $b_4 = 0.137$, we conclude that, averaging over the levels of the other factors, increasing temperature to 70°C improves yield. However, since $b_{2\cdot4} = -0.126$, the temperature effect may be influenced by Reaction time. At 15 min, the estimated temperature effect is $0.137 - 0.126(-1) = 0.263$, whereas at 30 min, the estimated temperature effect essentially disappears. A Time*Temperature interaction plot would display this, and would indicate a preference for the 15-min, 70°C combination.

2. Using a forward selection regression procedure with $\alpha = .05$ to select a hierarchical model for yield, we include two two-factor interactions, x_1x_4 and x_2x_4, and the three main effects x_1, x_2, and x_4. The analysis of variance for this fitted hierarchical model is as follows

Source	df	SS	MS	F
Model	5	0.8044	0.16088	6.690
Error	10	0.2405	0.02405	
Total (corrected)	15	1.0448		

What has changed from the previous analysis? We have dropped five terms from model (1.3) with hardly any decrease in the model sums of squares. The smaller MSE also results in a significant F statistic for the model (p-value = .0055) and smaller standard errors and smaller p-values for the estimated coefficients:

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	6.3819	0.0388	164.62	<.0001
x_1	−0.0956	0.0388	−2.47	.0333
x_2	−0.0106	0.0388	−0.27	.7896
x_4	0.1369	0.0388	3.53	.0054
x_1x_4	−0.0806	0.0388	−2.08	.0642
x_2x_4	−0.1256	0.0388	−3.24	.0089

Now, three or four effects appear to be statistically significant. With 10 df for error, twice the error degrees of freedom for Method 1, one might

presume that MSE = 0.024 provides a better estimate for σ^2. However, it is also possible that this MSE is smaller because we have overfit the model by including terms that have larger estimates just by chance.

3. Fit a saturated model to the ±1 coded factors and use the many coefficient estimates near zero to estimate σ. A Pareto plot of the 15 estimates is given in Figure 2.1. Lenth's (1989) procedure for estimating $\sigma/N^{1/2}$, the standard error of these estimates, is as follows:
   - Determine the median absolute estimate for the main effects and interactions from a saturated model and compute s_0 as 1.5 times this median. Here, $s_0 = 1.5(0.038125) = 0.0572$.
   - Exclude all estimates that exceed $2.5s_0$ in magnitude and recompute the median. Here, no estimates exceed $2.5s_0 = 0.143$, so the median remains 0.038125.
   - Compute PSE = 1.5 × median (of estimates less than $2.5s_0$). Here, PSE = 0.0572.

   Lenth's pseudo-standard-error (PSE) is an estimator for the standard error of the coefficients. Note how much larger it is than the standard error of 0.0388 from Method 2 above. Lenth's method provides a reasonable estimate for $\sigma/N^{1/2}$, provided only a few coefficients differ from zero. If this assumption is not correct, then Lenth's PSE will tend to overestimate $\sigma/N^{1/2}$. Lenth's PSE = 0.0572 corresponds to an estimate for σ of PSE$(N^{1/2}) = 0.0572(16^{1/2}) = 0.2288$. [Haaland and O'Connell (1995) show that the PSE is slightly biased upward when m is small, but the bias is only about 1% for $m = 15$.]

Term	Estimate
x4	0.1368750
x2*x4	-0.1256250
x1	-0.0956250
x1*x4	-0.0806250
x1*x2*x3	0.0743750
x2*x3*x4	0.0618750
x1*x2*x4	-0.0506250
x3	-0.0381250
x2*x3	-0.0331250
x1*x3	0.0168750
x3*x4	-0.0131250
x2	-0.0106250
x1*x2*x3*x4	0.0093750
x1*x3*x4	-0.0031250
x1*x2	-0.0006250

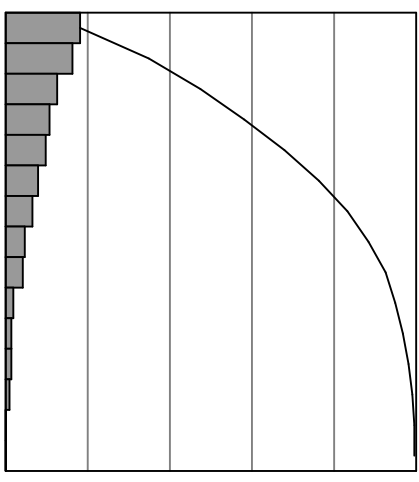
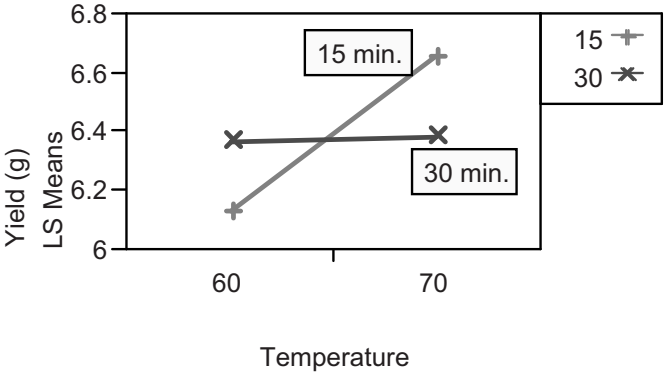


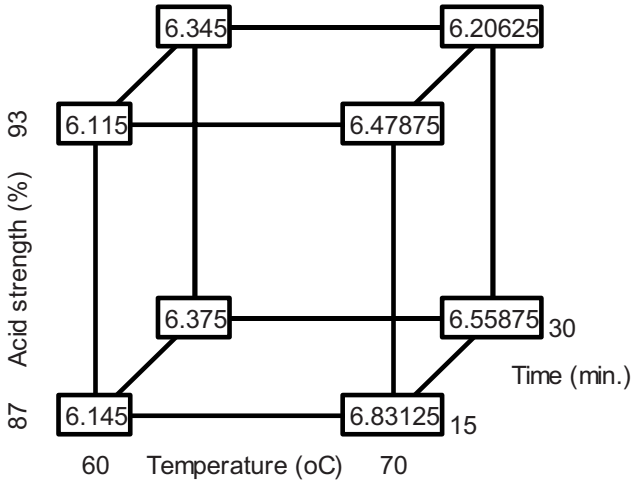
Fig. 2.1. Pareto plot of estimates from a saturated model for Davies's 2^4 experiment

These methods produced three different estimates for the error variance, ranging from Method 2's 0.024 to Method 3's $N(\text{PSE})^2 = 0.052$. Which coefficient estimates are statistically significant also varies from method to method. Which fitted model is best and which estimate is closest to the true σ^2 are unknown. For now, we discuss the possible interaction terms and then return to the discussion about estimators for σ^2.

With Method 1, $x_2 * x_4$ is the only statistically significant interaction. From its interaction plot



we conclude that 15 min at 70°C is preferable. With Method 2, we include an additional term or two that involve acid concentration (x_1). The model with 5 df yields the following cube plot for predicted yield:



If this model is correct, both low acid strength and shorter time are best when the process is run at 70°C. Note, however, that the predicted yield of 6.83 seems too optimistic, since no runs performed this well. If one were searching for still greater yields, then it seems reasonable to shift the experimental region in this direction and to experiment further with these factors.

It is sometimes the case that these methods differ even more in both their estimate for σ^2 and in the number of terms that are statistically significant. Choosing models without regard for statistical significance will surely lead to MSEs that underestimate σ^2. (See blunder-to-avoid #3 in Section 14.7.) To lessen the possible downward bias estimating σ^2 using a reduced model's MSE, we adopt the following conventions:

- Restrict final models to be hierarchical. For the isatin data, the MSE for Method 2 would have been even smaller than 0.024 if x_2 had been excluded due to its large p-value.
- Always include main effect terms when analyzing full factorial designs. Daniel (1959, p. 317) offered the following advice for constraining the use of negligible terms as part of an error variance estimate:

  > Nominate all effects, interactions and block contrasts[1] that are thought likely to be important *before* the experiment is completed. The corresponding contrasts are then to be excluded from further arbitrary allocation to error... Only those not nominated will be studied for possible allocation to error.

  Presumably prior to the isatin experiment, all four main effects were considered somewhat likely to be important. If so, then the variation explained by the x_3 term should not be pooled with error.

Fitting a model with all main effects ensures that practitioners are not misled by computer output reporting lack-of-fit tests when there is no replication. For the isatin model chosen under Method 2, some software would partition the error variation and report a lack-of-fit test with 2 df for lack-of-fit and 8 df for pure error. Since there is no replication, there can be no pure error. However, when a model with only three factors is fit to a 2^4 experiment, the software views the data as a replicated 2^3. Such confusion is avoided if one always includes the main effects. Further, retaining all the main effects in the model documents explicitly the relative unimportance of factors with negligible coefficients. Alternatively, JMP allows the user to designate a factor as "excluded," so that although not appearing in the model, it is recognized as a factor in the experiment. By this feature we may avoid spurious lack-of-fit tests.

With Methods 1 and 2, the MSE is used to estimate the error variance, and so the Student's t distribution is used to compute p-values for tests of individual coefficients. How to conduct tests based on Lenth's PSE will be addressed in the next subsection.

[1]The term "contrast" refers to a linear combination of the observations for which the coefficients sum to zero; i.e., the sum $\sum_{i=1}^{N} c_i y_i$ is a contrast if $\sum_{i=1}^{N} c_i = 0$. All main effect and interaction columns correspond to contrasts; see Table 1.4.

2.4.2 Tests for effects using Lenth's PSE

The previous subsection introduced the use of Lenth's PSE as a means of estimating the error variance without any replication, provided a majority of the true coefficients are zero. The steps in computing PSE are as follows:

- Determine s_0, 1.5 times the median absolute estimate from a saturated model fit to the ± 1 coded factors.
- Exclude any estimates that exceed $2.5s_0$ and recompute the median.
- Compute PSE $= 1.5 \times$ median (of estimates less than $2.5s_0$ in magnitude).

The logic behind this estimator is as follows. Suppose no effects are present so that $\mathrm{E}(b_i) = 0$ and $\mathrm{Var}(b_i) = \mathrm{E}(b_i^2) = \sigma^2/N$. Then one could use the average square of the b_i's to estimate σ^2/N. The median might also be used as an estimator that is more robust to outliers (i.e., to actual effects). Rather than using the median of the b_i^2's in an estimate for σ^2, Lenth (1989) proposed using the median of the $|b_i|$'s to estimate a multiple of σ. Since approximately half of a normal distribution with a mean of zero is between $-\sigma/1.5$ and $\sigma/1.5$, and the other half is further from the mean, s_0 is an initial rough estimate for the standard deviation of the b_i's. By excluding estimates that are more than $2.5s_0$ in magnitude, we eliminate estimates that appear to represent true effects. The remaining set of estimates is thus more nearly purged of estimates corresponding to β's that are not zero. Even if we compute the median from a list of estimates corresponding to effects, most of which are zero but with a few nonzero, the robustness of the median to outliers ensures that the PSE will not be greatly biased.

For cases with no error degrees of freedom, statistical software will often offer the option of computing Lenth t statistics as b_i/PSE. Percentiles of the sampling distribution of these statistics under the null hypothesis of no effects were estimated via simulation by Ye and Hamada (2000). The first part of Appendix C contains these IER critical values for Lenth t statistics. IER stands for "individual error rate," since these critical values (c_α^{IER}) are computed to limit the probability of a Type I error for each individual test across the set of tests. Occasionally in this book, we provide p-values, computed by simulation using JMP or the code in Appendix C, when analyzing unreplicated experiments via Lenth's procedure. For those wishing to conduct tests for a specified level α, simply use the IER critical values in Appendix C. Simulation is used to obtain critical values and p-values, since attempts at approximating the distribution of Lenth t statistics with a Student's t distribution have not achieved sufficient accuracy (Edwards and Mee 2008).

Consider again the example of Davies (1954). Table 2.7 gives the estimates for the saturated model, the PSE, Lenth t statistics, and p-values obtained by simulation. For software that does not furnish these p-values, an approximation for each p-value can be obtained using Appendix C. For instance, from the IER table in Appendix C we know that b_1, with Lenth $t = -1.672$ has p-value slightly above .10, since $c_{.10}^{\mathrm{IER}} = 1.701 > 1.672$.

Table 2.7. Estimates and p-values based on Lenth PSE for isatin data

Term	Estimate	PSE	Lenth t	p-Value
Intercept	6.3819	.0572		
x_1	−0.0956	0.0572	−1.672	.103
x_2	−0.0106	0.0572	−0.186	.861
x_3	−0.0381	0.0572	−0.667	.500
x_4	0.1369	0.0572	2.393	.037
$x_1 * x_2$	−0.0006	0.0572	−0.011	.992
$x_1 * x_3$	0.0169	0.0572	0.295	.783
$x_2 * x_3$	−0.0331	0.0572	−0.579	.598
$x_1 * x_4$	−0.0806	0.0572	−1.410	.160
$x_2 * x_4$	−0.1256	0.0572	−2.197	.048
$x_3 * x_4$	−0.0131	0.0572	−0.230	.828
$x_1 * x_2 * x_3$	0.0744	0.0572	1.301	.190
$x_1 * x_2 * x_4$	−0.0506	0.0572	−0.885	.354
$x_1 * x_3 * x_4$	−0.0031	0.0572	−0.055	.960
$x_2 * x_3 * x_4$	0.0619	0.0572	1.082	.265
$x_1 * x_2 * x_3 * x_4$	0.0094	0.0572	0.164	.876

To understand better the performance of Lenth's t statistics versus IER critical values, consider the case for $\alpha = .05$ and 15 estimates, where the critical value is $c_{.05}^{\text{IER}} = 2.156$; that is, we test all 15 estimates and declare ones larger in magnitude than 2.156(PSE) to be statistically significant. Since $\alpha = .05$ and $15(.05) = 0.75$, we expect, on average, 0.75 effects to be declared statistically significant, if in fact all true coefficients are zero.

To illustrate this, one million sets of 15 normal random variables with zero means were simulated. From each set, the PSE was calculated and the number of "estimates" found to exceed 2.156(PSE) was determined. The resulting distribution was as follows:

No. of Significant Effects Found	Frequency	Freq./10^6
0	604,881	0.6049
1	208,926	0.2089
2	94,398	0.0944
3	46,574	0.0466
4	23,907	0.0239
5	12,643	0.0126
6	6,130	0.0061
7	2,426	0.0024
8	94	0.0001
9	18	0.0000
10	3	0.0000

This distribution has a mean of 0.75, as required by using $\alpha = .05$ for 15 tests. Note that the risk of making one or more type I errors is $1 - 0.6049$

= 0.3951, or nearly 40%. This larger risk is called the experimentwise error rate (EER). It is informative to report both the individual error rate and the experimentwise error rate for a test procedure. Thus, for 2.156, the individual error rate is 0.05, whereas the experimentwise error rate is 0.395.

To control the experimentwise error rate, one may use the c_α^{EER} critical value table in Appendix C or those provided by Ye and Hamada (2000). These values were obtained by simulating $\max\{|b_1|, |b_2|, ..., |b_m|\}/\text{PSE}$ under the null hypothesis of no effects. For instance, for 15 contrasts and $\alpha = .10$, the EER critical value from Appendix C is $c_{.10}^{\text{EER}} = 3.505$. (Its individual error rate is about 0.011, and the expected number of Type I errors is $15(0.011) = 0.17$ when using the critical value 3.505.) In Table 2.7, no Lenth t statistics exceed 3.505; the largest is 2.393, which corresponds to an experimentwise error rate of 0.29 (i.e., if no true effects were present, nearly 30% of the time, one would obtain a largest Lenth t of 2.393 or greater). In experiments of this size, often controlling the individual error rate offers sufficient protection. However, when an experiment contains 2^6 or more treatment combinations, the number of eligible terms becomes so large that either controlling the experimentwise error rate or using a smaller α (e.g., .01) for IER is reasonable.

2.4.3 Alternatives to Lenth's *t* test

Hamada and Balakrishnan (1998) compared two dozen test procedures for unreplicated factorial designs. Most of these methods are intended to control the IER for each test. Lenth's method using IER critical values is one of the simplest, and it performs satisfactorily in terms of power. In Section 14.2, other more recent alternatives are discussed briefly.

Lenth's method, as originally proposed, is not the best for controlling the EER. A step-down version for Lenth's method proposed by Ye, Hamada, and Wu (2001) is certainly preferable. Section 14.2.1 illustrates this method and discusses some other alternatives for controlling the experimentwise error rate, including a simple step-up approach that utilizes standard F statistics for backward elimination regression. Section 14.2.2 makes the case that controlling EER is not usually of practical interest and argues for the intuitive alternative of controlling the proportion of Type I errors among all effects declared significant.

Finally, for any procedure such as Lenth (1989) based on the assumption of effect sparsity, be sure to fit a saturated model, since the method is based on the preponderance of negligible estimates. If one fits less than a saturated model, there will exist error degrees of freedom and software will use the MSE in constructing t-tests instead of the PSE, even if there is just 1 df for error. For 2^k factorial designs with no replication except at the center, most software will ignore the PSE. When the pure error degrees of freedom are very small and the sparsity of effects assumption is reasonable, then it is prudent to combine the MSE with the estimate for σ^2 that comes from Lenth's PSE. Section 14.1 presents two means for doing so.

2.5 Normal Plot of Effects and Other Analysis Tools

2.5.1 Normal and half-normal plot of effect estimates

Long before Lenth (1989) promoted the testing for effects in unreplicated experiments based on the sparsity of effects principle, Daniel (1959, 1976) and others urged that the effect estimates be plotted. If the sparsity of effects assumption is true, then for a 2^k factorial design, the majority of estimators for the coefficients in the saturated model (1.4) follow a normal distribution with mean 0 and variance $\sigma^2/2^k$. The $m = 2^k - 1$ estimates (excluding the intercept) are ordered from smallest to largest and plotted versus the standard normal quantiles Z_{P_i} $(i = 1, ..., m)$, where we use Blom's (1958) recommended proportions

$$P_i = (i - 0.375)/(m + 0.25). \tag{2.3}$$

For example, with $m = 15$, the largest estimate is plotted versus $Z_{14.625/15.25} = 1.739$ and the smallest estimate versus -1.739. In the normal plot of estimates, most are expected to fall along a line with an intercept of zero and (unknown) slope of $\sigma/N^{1/2}$, where, here, $N = 2^k$. For instance, Figure 2.2 shows the plot of the 15 estimated factorial effects from Table 2.7. The fitted line was constrained to have an intercept of zero and a slope equal to Lenth's PSE = 0.0572. The fact that a few of the 15 estimates fall below the line on the left and above the line on the right is weak evidence that these estimates correspond to effects that are present (i.e., $\beta_s \neq 0$). The closer the estimates fall along the line, the more consistent the data are with an assumption of no true effects.

Since the statistical significance of an estimate is generally based on its size $|b_s|$, a half-normal plot is seen as more useful than a normal plot by some (see Daniel 1959). For a half-normal plot of effects, we sort the absolute values of the estimates from smallest to largest and plot these versus the standard normal quantiles Z_{Q_i} $(i = 1, ..., m)$, where

$$Q_i = 0.5 + (i - 0.055)/(2m + 1.2). \tag{2.4}$$

Use of the proportions (2.4) was determine empirically and appears to be more accurate that Daniel's choice of $0.5+(i-0.5)/2m$. For a more accurate closed-form approximation of half-normal order statistic expected values, see Olguin and Fearn (1997, p. 460). A half-normal plot for the estimates in Table 2.7 is given in Figure 2.3. This plot reveals even more prominently the possibility of two or more active effects.

Statistical software such as JMP and Minitab automates the plotting of effects as in Figures 2.2 and 2.3, labeling the larger estimates. Such software may use different formula than (2.3) and (2.4), resulting in slight differences in the appearance of the plots. For example, JMP 7.0 uses $P_i = i/(m+1)$, which results in less extreme Z_{P_i}, whereas MINITAB 14 provides several options, with $P_i = (i - 0.3)/(m + 0.4)$ as the default.

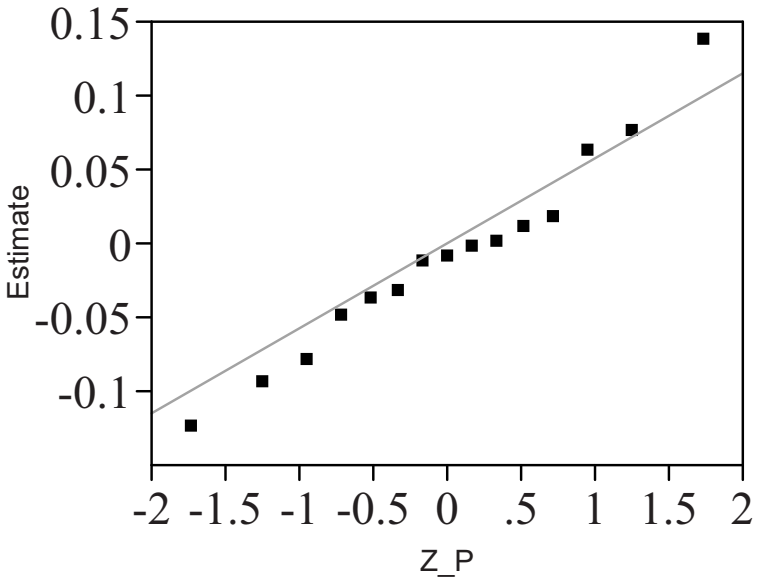


Fig. 2.2. Normal plot of effect estimates in isatin experiment

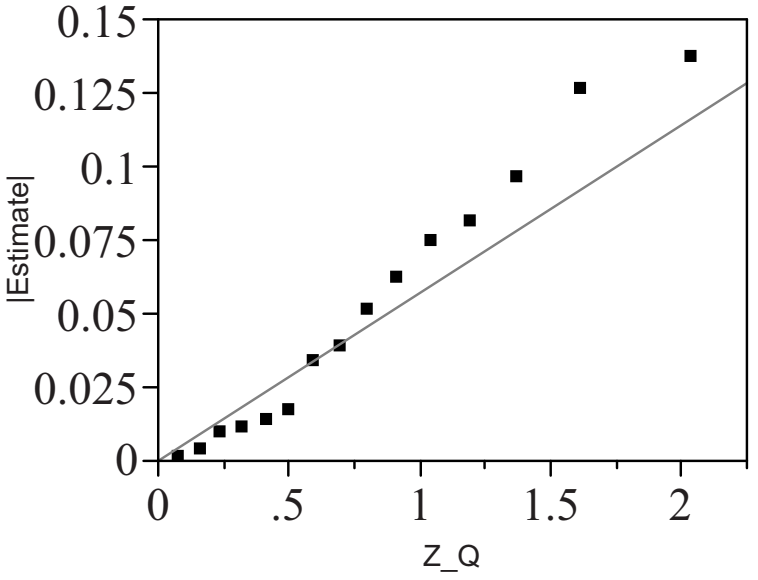
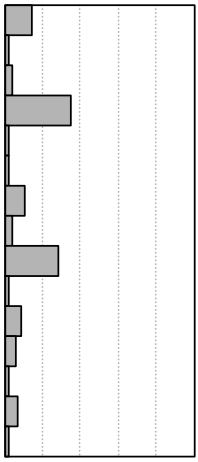


Fig. 2.3. Half-normal plot of effect estimates in isatin experiment

2.5.2 Bayesian analysis of 2^k factorial designs

Box and Meyer (1986) proposed an intuitive method for analyzing unreplicated 2^k experiments that does not involve tests of significance. Rather, it supposes that some fraction $1-\alpha$ of the effects are zero and that the remaining proportion α come from a normal distribution with a variance that is substantially larger than σ^2/N, the error variance for the estimates. A Bayesian framework combines these prior assumptions with the data and produces a "posterior" probability for each effect that it comes from the subset of active effects. For the isatin 2^4 experiment, the posterior probabilities as computed by JMP 7.0 are displayed in Figure 2.4.

Term	Estimate	Prior	Posterior
x_1	-1.672	0.2000	0.145
x_2	-0.186	0.2000	0.025
x_3	-0.667	0.2000	0.045
x_4	2.393	0.2000	0.353
x_1*x_2	-0.011	0.2000	0.024
x_1*x_3	0.295	0.2000	0.027
x_1*x_4	-1.410	0.2000	0.103
x_2*x_3	-0.579	0.2000	0.040
x_2*x_4	-2.197	0.2000	0.284
x_3*x_4	-0.230	0.2000	0.026
x_1*x_2*x_3	1.301	0.2000	0.090
x_1*x_2*x_4	-0.885	0.2000	0.056
x_1*x_3*x_4	-0.055	0.2000	0.024
x_2*x_3*x_4	1.082	0.2000	0.070
x_1*x_2*x_3*x_4	0.164	0.2000	0.025



Posterior Prob that sample is uncontaminated: 0.3395

Fig. 2.4. Posterior probabilities for effects in isatin experiment; $K = 10$, $\alpha = .20$

Note that the estimates reported by JMP in Figure 2.4 are actually Lenth t statistics (refer to Table 2.7). For 13 of the 15 estimates, the posterior probability is lower than the prior probability of .2. The larger the magnitude of the estimate, the larger the posterior probability. However, even the largest estimate ($b_4 = 0.13$ with Lenth $t = 2.393$) only has a posterior probability of .35. Thus, although there is some evidence for one or two effects, that evidence is not compelling.

Recall that using an individual error rate of 0.05, the two largest estimates are statistically significant, whereas controlling the EER at 0.25 or smaller, no effects are statistically significant. The posterior probabilities in Figure 2.4 were obtained assuming 20% of effects are active and that they have a variance 10 times σ^2/N (since JMP's $K = 10$). Since the largest estimates

were not much larger than the others, even the largest estimates are deemed more likely to correspond to null effects. However, if we lower K to 5, the posterior probabilities for b_4 and $b_{2\cdot4}$ increase to 0.59 and 0.51, respectively. The conclusion is still that the evidence for these effects being active is rather weak, given a prior expectation that $3/15 = 20\%$ of the effects would be active. For more details on the computations, see Box and Meyer (1986, 1993). For a comparison of the Bayesian approach with Lenth's method, see Haaland and O'Connell (1995).

2.6 Diagnostics for a Fitted Model

The residual e_i is the difference between the i^{th} observed response y_i and the corresponding predicted value $\hat{y}_i$ from a fitted model—that is, $e_i = y_i - \hat{y}_i$ ($i = 1, ..., N$). Residuals indicate the extent of disagreement between an assumed model and the actual data, and so provide a means of checking both the tentative model for $E(y)$ and the assumptions regarding ϵ.

2.6.1 Plotting residuals versus predicted y

Plotting e_i versus $\hat{y}_i$ is particularly helpful for assessing model adequacy, provided there are enough error degrees of freedom. (The error degrees of freedom indicate the amount of information in the residuals.) To illustrate this point, consider several residual plots for the Bouler et al. (1996) data discussed in Section 2.3. Figure 2.5a displays residuals versus predicted values for the full factorial model as summarized in the Table 2.4 ANOVA. Here, we have only 7 df for error: 1 df for lack-of-fit and 6 df for pure error. This residual plot is not very useful, since it simply shows the pure error variation among the centerpoint runs and the statistically significant lack-of-fit due to the centerpoint residuals being predominantly positive. (If one were to fit a saturated model by adding the term ${x_1}^2$, then the residuals for all the factorial points would be zero, and the residuals for the center runs would average zero. Plotting these residuals would have no value.)

Consider a second residual versus predicted plot based on a reduced model that eliminates all interactions involving x_3 or x_4. With 22 interactions removed, this model has 29 df for error, and its residuals are displayed in Figure 2.5b. This residual plot is more useful. First, it shows that the pure error variation at the center is small compared with the variation in other residuals, so that the lack-of-fit for this reduced model must be statistically significant. In addition, there is more variation in compressive strength when the expected compressive strength is above 30. Finally, residuals corresponding to the 10 observations with zero compressive strength have predicted values ranging from -3.4 to 6.7. Overall, this residual plot reflects an unsatisfactory model. In Section 2.7 we will discuss how using a transformation for y can improve the model fit in such occasions. For instance, if we fit the same model (with

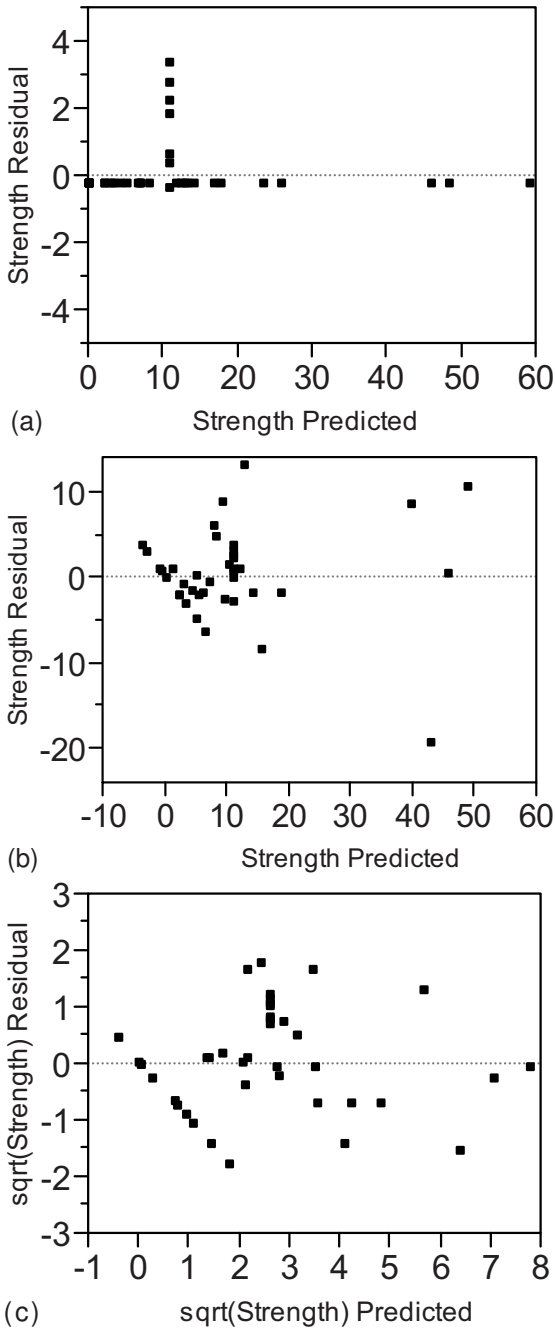


Fig. 2.5. Residual plots for (a) full factorial model for y = strength; (b) reduced model for y = strength; (c) reduced model for $y = (\text{strength})^{1/2}$

29 df for error) to the square root of compressive strength, as seen in Figure 2.5c, the variation in the residuals appears more consistent across the range of predicted values and only 1 of the 10 observations with $y_i = 0$ has a negative predicted value (-0.4).

Residual versus predicted plots are particularly helpful for spotting violations of the assumption of constant $\text{Var}(\epsilon)$. When there is replication at several or all of the treatment combinations, there exist tests for checking the equality of $\text{Var}(\epsilon)$. Common tests available in software include Bartlett's test for equality of variances and the more robust tests by O'Brien (1979, 1981).

2.6.2 Plotting residuals versus run order

Plotting residuals versus $\hat{y}_i$ is only one of several useful means for examining the residuals. When the data are time (or spatially) ordered, it is important to plot the residuals versus that order. Such a plot is displayed in Figure 11.2, where a possible shift in the measurement process is revealed. Autocorrelation of the errors is another possibility related to time-ordered experimental runs. Use the Durbin–Watson statistic to check for first-order autocorrelation. When applying this test using any software, be sure to have the data sorted by run order. The Durbin–Watson test is more important when one is experimenting with a highly variable process where such correlation is deemed likely. Although randomization of run order does not eliminate trend or autocorrelation for ϵ, it does offer protection against the effects of such problems in most situations (Box, Hunter, and Hunter 2005, pp. 404f). For further discussion, see Section 13.5.

2.6.3 Plotting residuals in normal quantile plot

When there is a large amount of data and R^2 is low, then the distribution of ϵ becomes important. In such cases, one may construct a normal quantile plot of the residuals. For instance, see Figure 4.9. By contrast, when R^2 is above 90%, the distribution of ϵ has minimal importance, since the distribution of the residuals will reflect lack-of-fit more than it will the actual distribution of ϵ. For this reason, we do not routinely construct a normal plot of residuals for examples in this book.

2.6.4 Identifying outliers using Studentized residuals

Spurious y_i values are a serious concern, especially for small, unreplicated experiments, because of their influence on the fitted model. However, an observation that appears to be an outlier under one model may appear reasonable under a different model. For instance, the large negative residual displayed in Figure 2.5b is problematic if the error variance is constant. However, if the error variation increases as strength increases, then the same observation no

longer appears extreme. The less data one has, the more ambiguity exists regarding how to interpret such runs. A simple, practical approach to handling suspected outliers is to fit models both with and without the runs, to see their impact on the conclusions. Daniel (1959, pp. 331f) pointed out that for two-level factorials, a single outlier will bias every effect estimate by (±) the same amount and that this will alter the half-normal plot of effect estimates to have no clump of estimates at zero. The case study in Section 4.2 will illustrate how to address the problem of more outliers.

Literature about outliers in regression is extensive. Beckman and Cook (1983, Section 4.2) provided an excellent overview; see also Gray and Woodall (1994). The Studentized residual is defined as

$$r_i = e_i/[(1 - h_{ii})\text{MSE}]^{1/2}, \tag{2.5}$$

where h_{ii} is the $(i,i)^{\text{th}}$ element of the "hat" matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. For a 2^k factorial (with $n \geq 1$ observations at each treatment combination and no centerpoint runs), $h_{ii} = r/N$, where r is the number of columns in X.

The distribution of the maximum Studentized residual can be simulated for any model matrix $\mathbf{X}$. Appendix D provides a simple simulation program that may be used to determine upper (10th and 5th) percentiles for the maximum (in absolute value) Studentized residual, and the probability of getting a maximum residual as large as that observed for a particular model. This provides a quick reference regarding whether any observations may be considered outliers. Gray and Woodall (1994) showed that the maximum value for (2.5) is $(N - r)^{1/2}$. When the degrees of freedom for error are 4 or less, there is no point checking for extreme outliers.

If the i^{th} observation is deleted and the same model fit to the data, the error sum of squares will decrease by $e_i^2/(1-h_{ii})$ and the i^{th} "deleted residual," the difference between y_i and the predicted value for the omitted observation, is

$$d_i = e_i/(1 - h_{ii}). \tag{2.6}$$

The Studentized version of (2.6) is the same as (2.5), except that the estimate for σ^2 is based on $\text{MSE}_{(-i)}$, the mean square error with the i^{th} observation excluded, which is

$$\text{MSE}_{(-i)} = [\text{SSE} - e_i^2/(1 - h_{ii})]/[N - r - 1], \tag{2.7}$$

where SSE is the error sum of squares, $\sum_j^N e_j^2$. The most convenient means for computing (2.6) is to add a dummy variable column to the $\mathbf{X}$ matrix with the value 1 for the i^{th} row and 0 otherwise. The regression coefficient for this column will equal the deleted residual (2.6), the MSE will equal (2.7) and the t statistic for the coefficient of the dummy column will equal the Studentized residual for the i^{th} observation. Typically, an observation attracted attention simply because it had one of the largest residuals. Based on the Bonferroni inequality, one may multiply the p-value by N to get the approximate probability that the biggest residual would be larger than this just by chance.

We now illustrate these computations with a small example to reinforce the concepts. Suppose we fit the isatin yield data from Table 2.6 with a model containing all four main effects and the $x_2{*}x_4$ interaction and consider whether any residual is unusually large. (Perhaps that is the reason we saw few significant terms.) The largest residual is for the fifth observation, with $y_5 = 6.12$ and $\hat{y}_5 = 6.4175$. With MSE $= 0.3212/10$,

$$e_5 = y_5 - \hat{y}_5 = 6.12 - 6.4175 = -0.2975,$$

$$r_5 = e_5/[(1 - h_{55})\text{MSE}]^{1/2} = -0.2975/[(1 - 6/16)0.03212]^{1/2} = -2.0997.$$

This is not unusually large. Using the simulation program in Appendix D, we determine that there is a 39.5% chance of getting a Studentized residual this far from zero.

If one were to delete the fifth observation and refit the model, the predicted value for this observation is 6.596 and

$$d_5 = 6.12 - 6.596 = -0.2975/(1 - 6/16) = -0.476.$$

If instead of deleting this observation, one adds a dummy variable for the fifth observation, the estimated model becomes

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	0.412	0.0370	11.113	.000
x_1	−0.125	0.0370	−3.385	.008
x_2	0.019	0.0370	0.516	.618
x_3	−0.008	0.0370	−0.226	.826
x_4	0.107	0.0370	2.892	.018
$x_2{*}x_4$	−0.155	0.0370	−4.195	.002
Dummy_5	−0.476	0.1787	−2.664	.026

with mean square error

$$\begin{aligned}\text{MSE}_{(-5)} &= [0.3212 - (-0.2975)^2/(1 - 6/16)]/[16 - 6 - 1] \\ &= 0.1796/9 = 0.0200.\end{aligned}$$

The standardized deleted residual for y_5 is −2.664 (p-value = .026). However, a p-value as small as .026 is typical for the most extreme outlier. Multiplying by $N = 16$, we obtain $16(.026) = 0.414$; this Bonferroni upper bound is only slightly larger than the exact probability of .395 found using the Appendix D simulation. Assuming that this model is correct, there is no indication of any outliers among these data.

In Chapter 4, we analyze case studies in which many outliers will be evident.

2.7 Transformations of the Response

Example 2.3. Drill Advance Rate for 2^4

Daniel (1976) introduced the use of transformations for y in a section titled "Looking for Simple Models." His 2^4 example involving the advance rate of a stone drill illustrates clearly the potential advantages. The data are displayed in Figure 2.6.

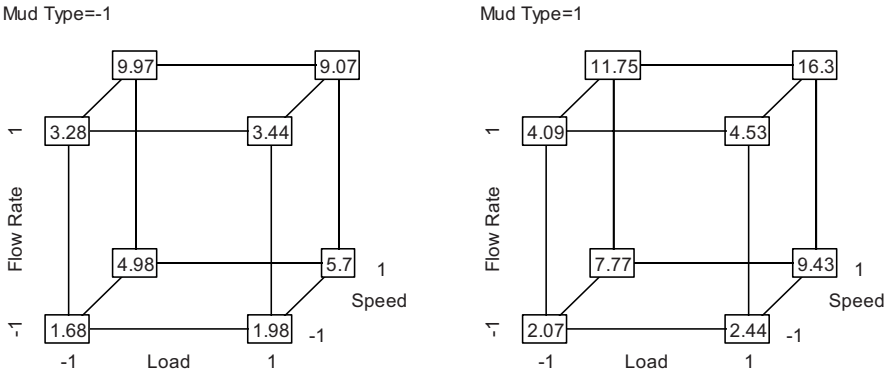


Fig. 2.6. Cube plot of Example 2.3 advance rate data from Daniel (1976)

Fitting a saturated model to these data, we obtain a half-normal plot of the effects (see Figure 2.7). This plot is pleasing, in that three of the main effects have statistically significant estimates, based on their Lenth t statistics. A reduced model would certainly contain these three terms and possibly the $x_{\text{Speed}} * x_{\text{Flow}}$ and $x_{\text{Speed}} * x_{\text{Mud}}$ interactions, since their estimates also stand off the line. The resulting model,

$$\begin{aligned}\hat{y} = \; & 6.15 + 1.65x_{\text{Flow}} + 3.22x_{\text{Speed}} + 1.14x_{\text{Mud}} \\ & +0.75x_{\text{Flow}} * x_{\text{Speed}} + 0.80x_{\text{Flow}} * x_{\text{Mud}},\end{aligned} \tag{2.8}$$

explains 95% of the variation in advance rate. However, the normal plot of effects for the saturated model (Figure 2.8) looks peculiar in that all 15 estimates are positive, so that the estimates are far from the line through the origin; there is no clump of estimates centered about 0. In addition, the residuals from the reduced model (2.8) are more scattered at large predicted advance rates (see Figure 2.9). All of these plots indicate that we are missing some systematic variation with our model, even though $R^2 = .95$.

Daniel (1976) fitted models for nine different transformations of y = advance rate, including different powers of y, and the log transformations

$$\ln(y + c) \tag{2.9}$$

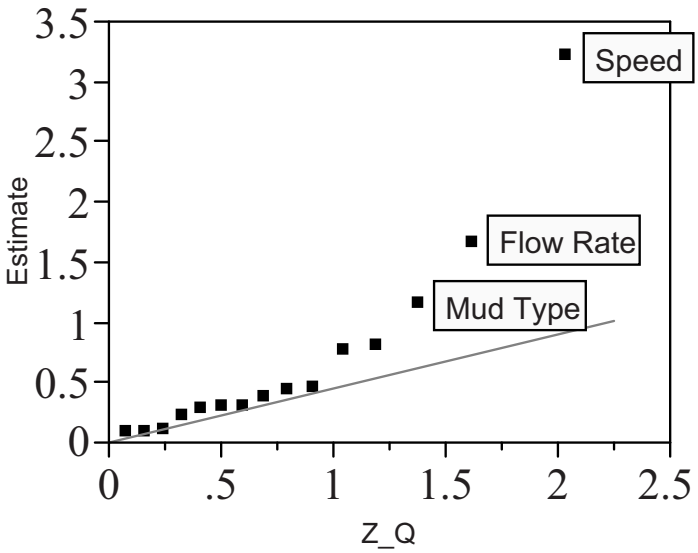


Fig. 2.7. Half-normal plot of effects for Daniel's drill data

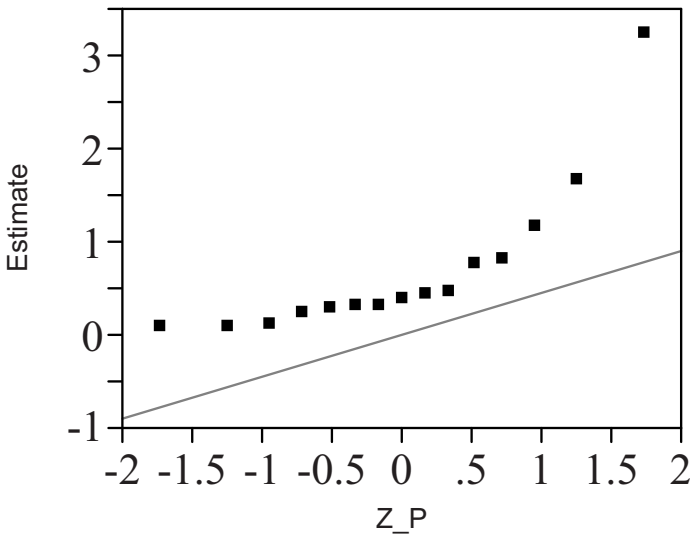


Fig. 2.8. Normal plot of effects for Daniel's drill 2^4

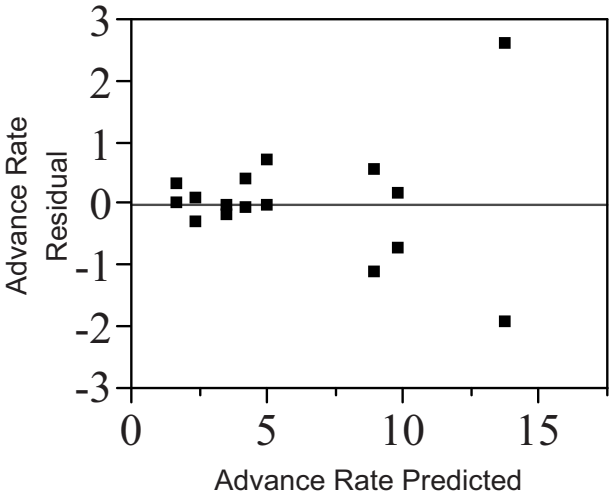


Fig. 2.9. Residuals versus predicted advance rate for reduced model (2.5) for Daniel's drill 2^4

with different constant shifts c. The family of transformations (2.9) is valid, provided $c > -\min\{y_1, ..., y_N\}$. The most popular set of transformations today is the family proposed by Box and Cox (1964):

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/(\lambda \dot{y}^{\lambda-1}), & \text{if } \lambda \neq 0 \\ \ln(y)\ \dot{y}, & \text{if } \lambda = 0 \end{cases} \tag{2.10}$$

where $\dot{y}$ is the geometric mean $\dot{y} = \prod_{i=1}^{N} y_i^{1/N}$. By normalizing the power transformation as in (2.10), the value of λ for which the error sum of squares is minimized is the maximum likelihood estimator for λ. Since the normalized transformation nearly makes the total sum of squares for $y^{(\lambda)}$ invariant to λ, the λ that minimizes the SSE essentially maximizes R^2.

Suppose we fit an additive model in the four factors for advance rate. A plot of the error sum of squares for different transformations $-2 \leq \lambda \leq 2$ as produced by JMP is shown in Figure 2.10. JMP searches the grid $\{-2, -1.8, -1.6, ..., 2\}$ and determines that transformed values corresponding to $\lambda = 0$ have the smallest error sum of squares for the additive model.

The actual maximum likelihood estimator here is $\hat{\lambda} = -0.05$, but taking $\lambda = 0$ is simpler and produces essentially the same result. A 95% confidence interval for λ is the interval of values that produce an error sum of squares below the horizontal line in Figure 2.10. For details on the computation, see Box and Cox (1964) or Montgomery and Peck (1992). In Figure 2.10 this confidence interval is narrow for two reasons. First, the ratio maximum$\{y_i\}$/minimum$\{y_i\}$ = 9.7. When this ratio is less than 2, nonlinear transformations will have lit-

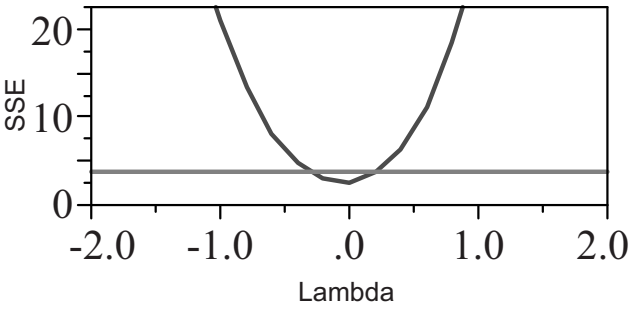


Fig. 2.10. Error sum of squares from additive model for different transformations of advance rate, $-2 \leq \lambda \leq 2$

tle effect on the result, and we will be indifferent to models for a wide range of λ. By contrast, when the maximum is an order of magnitude larger than the minimum, nonlinear transformations have a pronounced effect, and so some transformations are clearly better than others. Second, we considered different transformations for the additive model, which leaves much of the variation unexplained. If one were to choose a model with more terms, then many different λ's may explain most all the variation, and so again the choice for the best λ will not be so clearly indicated. For instance, if the Box–Cox transformation is applied fitting the two-factor interaction model to Daniel's data, the confidence interval for λ is $(-1.23, 0.05)$; from this fit, either the log or reciprocal transformation is acceptable. We prefer the log transformation because the histogram for $\ln(y_i)$ is less skewed than the histogram for $1/y_i$. In addition, the resulting model matches the engineering expectation that the factor effects might be multiplicative.

The fitted additive model for predicted ln(advance rate) is:

$$\widehat{\ln(y)} = 1.600 + 0.065x_{\text{Load}} + 0.290x_{\text{Flow}} + 0.577x_{\text{Speed}} + 0.163x_{\text{Mud}}. \quad (2.11)$$

Taking the exponential of (2.11) produces an estimate for the median (not the mean) advance rate:

$$e^{\widehat{\ln(y)}} = 4.953(1.067)^{x_{\text{Load}}}(1.336)^{x_{\text{Flow}}}(1.781)^{x_{\text{Speed}}}(1.177)^{x_{\text{Mud}}}$$

since $e^{1.6} = 4.953$, $e^{.065} = 1.067$, etc. The predicted median rates range from 1.66 to 14.8.

In Daniel's drill example, the simple additive model in ln(advance rate) accounted for 98.5% of the variation, whereas modeling advance rate directly would have required a model with many terms to achieve an R^2 so large. There are additional reasons for considering transformations. First, if there is substantial error variation and the variance is not constant, then ordinary least squares estimation loses efficiency. When the error variation depends on

$E(y)$, then choosing a suitable function f and modeling $f(y)$ rather than y directly can resolve the unequal variance problem and keep the estimation simple. This is the case for the Bouler et al. (1996) data; recall the improved residual plot in Figure 2.5c for $y = (\text{strength})^{1/2}$.

In this section, we have addressed applications in which the choice of a transformation $f(y)$ is determined empirically. Sometimes the nature of the response y suggests what transformation is appropriate. For example, when y is a count, following a Binomial or Poisson distribution, known transformations will stabilize the variance (see Sections 2.8.1 and 2.8.2). Another common response is the standard deviation. Section 2.8.3 details why the log transformation is appropriate for variances and standard deviations.

2.8 Analysis of Counts, Variances, and Other Statistics

For some experiments, responses are counts. For example,

- number of flaws in a door panel
- number of respondents to an email solicitation
- number of defective parts in a sample of 20

Counting the number of good (or bad) parts is not as informative as collecting quantitative data on each part. For example, it is better to measure the breaking strength on each of a sample of parts than it is to simply know how many failed at a certain stress. However, in some applications, quantitative data are either too expensive or impossible to collect and count data are all that are available. Count data routinely violate the assumption of constant variance for ϵ, and so specialized methods are required. The simplest of these methods is to use least squares for a transformation of the response. When the sample sizes at each treatment combination are large, use of least squares is often justified. For cases where sample sizes are smaller, other methods are recommended. After discussing and illustrating the options for count data, we discuss the common case of modeling a variance and then briefly mention analyzing correlations, ratios, lifetimes, directions, and functional responses.

2.8.1 Modeling Binomial proportions

When the measured outcome at a treatment combination is the proportion of n trials having a characteristic of interest, the Binomial distribution is generally appropriate. Let c denote the number of cases having the characteristic of interest and let $\hat{p} = c/n$ denote the observed proportion. If the outcomes of the individual trials are independently distributed and the number of trials is fixed, then c has a Binomial distribution with parameters n and p, where $p = E[\hat{p}]$. There are two problems associated with modeling $\hat{p}$. First, since the variance of $\hat{p}$ depends on p, any factor that affects the mean also affects the

variance. Thus, the typical constant variance assumption will be violated. Second, since $0 \le p \le 1$, a fitted model for p may result in predicted proportions outside the feasible range.

$\mathrm{Var}(\hat{p}) = p(1-p)/n$ is maximum at $p = .5$ and is relatively stable over the interval $.3 \le p \le .7$. However, for problems where the proportions are not confined to this range and where least squares estimation is to be used, it is best to model some function of $\hat{p}$ that stabilizes the variance. One option is

$$f_a(\hat{p}) = \arcsin(\sqrt{\hat{p}}).$$

Figure 2.11 shows how this function is essentially linear over the range $.3 \le \hat{p} \le .7$, but it accentuates differences among more extreme values for $\hat{p}$, where $\hat{p}$ is less variable.

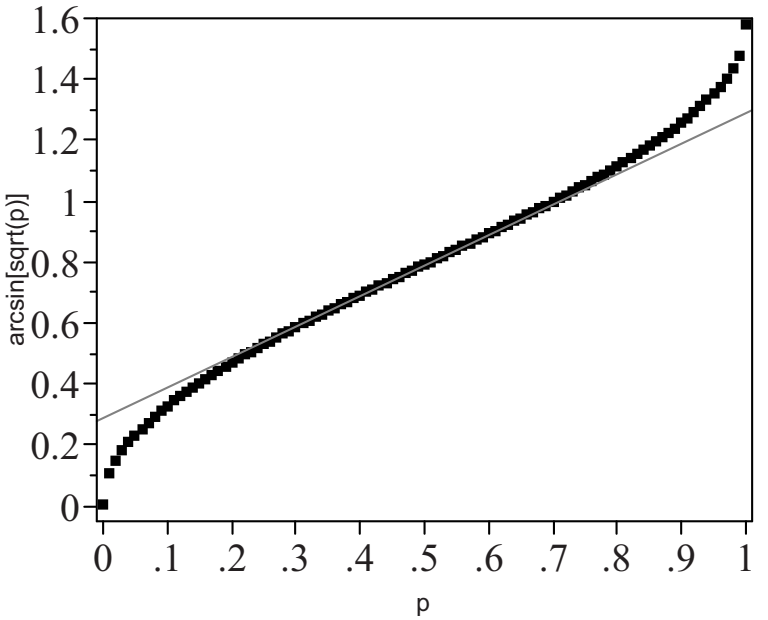


Fig. 2.11. $f_a(p) = \arcsin(\sqrt{p})$ transformation; slope ≈ 1 for $.3 \le p \le .7$

Freeman and Tukey (1950) recommended a modification to the transformation $f_a(\hat{p})$. The Freeman–Tukey transformation for Binomial proportions is

$$\begin{aligned} f_{\mathrm{FT}}(\hat{p}) &= f_a[\hat{p}n/(n+1)] + f_a[(\hat{p}n+1)/(n+1)] \\ &= \arcsin[\sqrt{c/(n+1)}] + \arcsin[\sqrt{(c+1)/(n+1)}], \end{aligned} \tag{2.12}$$

where c is the number of cases out of n with the characteristic of interest. Note that this transformation depends on both $\hat{p}$ (or c) and n.

Figure 2.12 shows the variance of $f_a(\hat{p})$ and $f_{\rm FT}(\hat{p})$ for $n = 10$ and Figure 2.13 shows the same for $n = 40$. Reference lines are drawn at $1/n$ and $1-1/n$ in each figure. In Figure 2.12, $\mathrm{Var}(\hat{p}) = p(1-p)/n$ is also displayed. For $n = 10$, over the interval $.1 \le p \le .9$, $\mathrm{Var}(\hat{p})$ ranges from .009 to .025, a max/min ratio of 2.78. Both variance-stabilizing transformations do much better. For $f_a(\hat{p})$, the max/min ratio is $= 0.04202/0.02857 = 1.45$, whereas the Freeman–Tukey transformation has the max/min ratio of $0.09998/0.0913 = 1.095$. Freeman and Tukey (1950) stated that (2.12) produces variances within $\pm 6\%$ of $1/(n+0.5)$ for almost all cases where the expected proportion p is between $1/n$ and $1-1/n$. This corresponds to data where the expected count is at least 1 and not more than $n-1$. For $n = 40$, the max/min ratio for the variance over the interval $.025 \le p \le .975$ is 1.61 for $f_a(\hat{p})$ and 1.12 for $f_{\rm FT}(\hat{p})$. Use of (2.12) as the response is recommended provided one has few sample proportions of 0 or 1. Given a fitted model for (2.12), the inverse transformation (Miller 1978) is

$$\hat{p}(f) = 0.5\{1 - \mathrm{sgn}(\cos f)[1 - (\sin f + (\sin f - 1/\sin f)/n)^2]^{1/2}\},$$

where f is the predicted value for $f_{\rm FT}$ and $\mathrm{sgn}(\cos f)$ denotes the sign of $\cos f$.

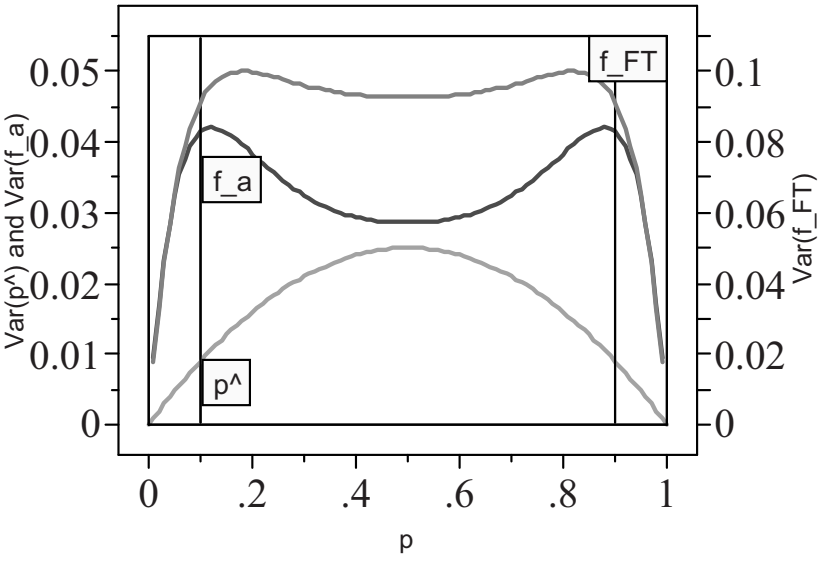


Fig. 2.12. Variance for $\hat{p}$, $f_a(\hat{p})$, and $f_{\rm FT}(\hat{p})$ for $n = 10$

Arber et al. (2006) conducted a 2^4 factorial experiment to see how gender, age, race, and social class affected physicians' diagnoses and follow-up recommendations for simulated coronary heart disease patients. $N = 256$ physicians

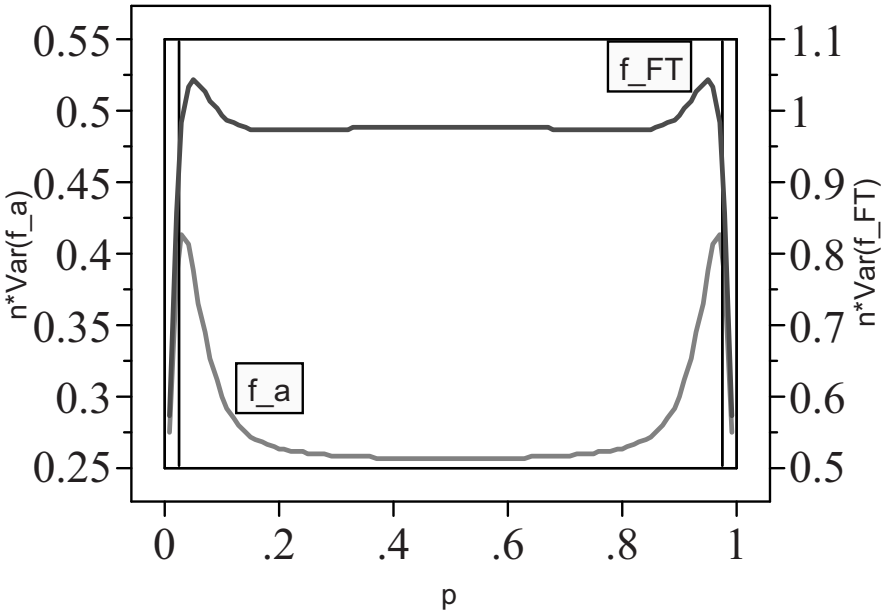


Fig. 2.13. n·Variance for $f_a(\hat{p})$ and $f_{\mathrm{FT}}(\hat{p})$ for $n = 40$

took part, with 16 assigned to each treatment combination, 8 from the United States and 8 from the United Kingdom. Thus, for analyses that ignore country, $n = 16$ per treatment combination, and the Freeman–Tukey transformation (2.9) will effectively stabilize the variance at $1/16.5$ for p in the interval (.067, .933). For analyses of individual country data, $n = 8$, and so the range of p for which the variance of the Freeman–Tukey transformed proportions is near $1/8.5$ is constrained to (0.125, 0.875). Since several of the characteristics of interest occurred for over 90% or less than 10% of the doctors, the country-specific data cannot be effectively analyzed using least squares and the response $f_{\mathrm{FT}}(\hat{p})$.

When some expected counts are close to zero or n, a linear model for p or $f_{\mathrm{FT}}(p)$ may not be suitable. An alternative is to model the log-odds, $\ln[p/(1-p)]$, with estimation via maximum likelihood rather than least squares. Two advantages of modeling the log-odds are (i) any predicted value for the log-odds corresponds to a value for p within the interval (0,1) and (ii) the models are meaningful to interpret. For instance, the additive model (1.3) translates into a model of independence, whereas models with some interactions are interpretable in terms of conditional independence. Although maximum likelihood estimation may require iteration, the required software is widely available. For a useful reference, see Collett (2002).

2.8.2 Modeling Poisson count data

Example 2.4: 2^4 factorial with y = number of blemishes

Hsieh and Goodwin (1986) described an experiment to reduce the number imperfections in a grille used by a Chrysler assembly plant. Porosity problems caused the blemishes, and a 16-run experiment was performed in search of a remedy. Four factors were Mold pressure (x_1), Priming method (x_2), Thickening process (x_3), and Viscosity (x_4). Pressure and Viscosity are quantitative factors, although the actual levels used were not reported; the other two factors are qualitative. The 16 treatment combinations in the order listed by the authors are shown in Table 2.8, along with the total number of "pop" defects observed for each. The observed counts range from 3 to 99 pops. We are not told whether a single part or multiple parts were inspected at each treatment combination.

Table 2.8. Hsieh and Goodwin (1986) experiment

x_1	x_2	x_3	x_4	Total No. Pops, c
−1	−1	1	−1	66
−1	−1	−1	−1	19
−1	1	1	1	3
−1	1	−1	1	7
−1	1	−1	−1	4
−1	1	1	−1	17
−1	−1	−1	1	99
−1	−1	1	1	5
1	1	1	−1	4
1	1	−1	−1	3
1	−1	1	1	5
1	−1	−1	1	14
1	−1	−1	−1	7
1	−1	1	−1	14
1	1	−1	1	5
1	1	1	1	8

If flaws arise individually and independently, then the data will follow a Poisson distribution; refer to any probability book for details (e.g., Ross 1998). Modeling Poisson counts c using ordinary least squares is not recommended, since any factor that affects the mean also affects the variance. For the Poisson distribution, the mean and variance are equal. Thus, if the factors do affect the mean, then the assumption of constant variance will be violated. Several alternative approaches are more appropriate:

- Use the simple, variance-stabilizing transformation $\sqrt{c}$. Think of this transformation as taking each observation c and dividing by an estimate of its

standard deviation, $\sqrt{c}$. If we divided c by its true standard deviation, $\sqrt{E(c)}$, the resulting standardized variable $c/\sqrt{E(c)}$ would have a variance of 1, whatever $E(c)$ is. Because the numerator and denominator of $c/\sqrt{c}$ are correlated, $\sqrt{c}$ has a variance smaller than 1 but one that is insensitive to $E(c)$.

- Use the Freeman and Tukey (1950) transformation for Poisson counts, $\mathrm{FT}(c) = (\sqrt{c} + \sqrt{c+1})/2$.
- Model c directly, using an estimation method other than least squares—for example, weighted least squares or maximum likelihood of a generalized linear model (GLM) (see Wu and Hamada 2000, p. 568). However, modeling c directly causes the coefficient estimators to be correlated, due to the nonconstant variance.

Figure 2.14 displays the variance of $\sqrt{c}$, $(\sqrt{c} + \sqrt{c+1})/2$, and c on the same plot as a function of $E(c)$. The right axis labels values for $\mathrm{Var}(c)$, and the straight line $y = x$ indicates the equality of $E(c)$ and $\mathrm{Var}(c)$. The left axis denotes the variance for both transformations of c. The curve with the smaller peak near 0.4 is $\mathrm{Var}(\sqrt{c})$ and the curve with the peak of 0.5 for $E(c) = 1$ is for the Freeman–Tukey transformation. The Freeman–Tukey transformation is essentially perfect for stabilizing the variance if $E(c) \geq 5$, but it is slightly worse than $\sqrt{c}$ if some expected counts are below 2.5.

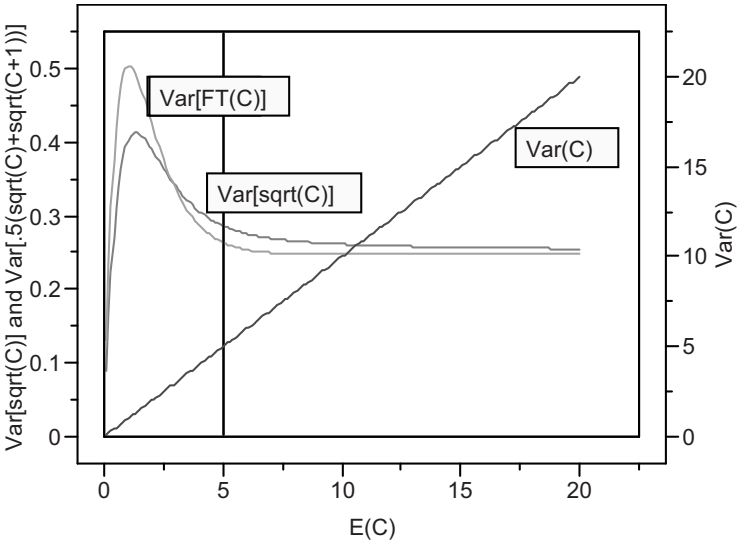


Fig. 2.14. Sqrt(Poisson count) as a variance-stabilizing transformation when expected count ≥ 5

For the Hsieh and Goodwin data, with one-fourth of the data being 3 or 4, we opt for the simpler square root transformation, since the Freeman–Tukey transformation is not better at stabilizing the variance for expected counts of 3 or less. Fitting a full factorial model for $y = (\text{number of pops})^{1/2}$, we obtain the 15 least squares estimates and display these in a half-normal plot (see Figure 2.15). In addition to drawing a line through the origin with a slope equal to Lenth's PSE (= 0.517), we draw a second line with a slope of $[0.3/16]^{1/2} = 0.137$, which would be the approximate standard error of the least squares estimates if in fact the data followed a Poisson distribution. (The variance of 0.3 in this calculation is taken from Figure 2.14.) The discrepancy between these two lines indicates that either the sparsity of effects assumption is violated, making Lenth's PSE too large, or the actual standard error is much larger than 0.137 because the "pop" defects do not follow a Poisson distribution. We suspect the latter, since otherwise a saturated model would be required to account for the observed data. The largest two estimates (in magnitude) are $b_2 = -1.11$ and $b_{3.4} = -1.05$, both with Lenth t statistics exceeding 2. This is evidence that the priming method coded "+1" is preferred, and that which thickening process is better depends on the viscosity factor's level. The third largest effect is for mold pressure ($b_1 = -0.88$); although not statistically significant, it suggested to Hsieh and Goodwin that the higher pressure is better.

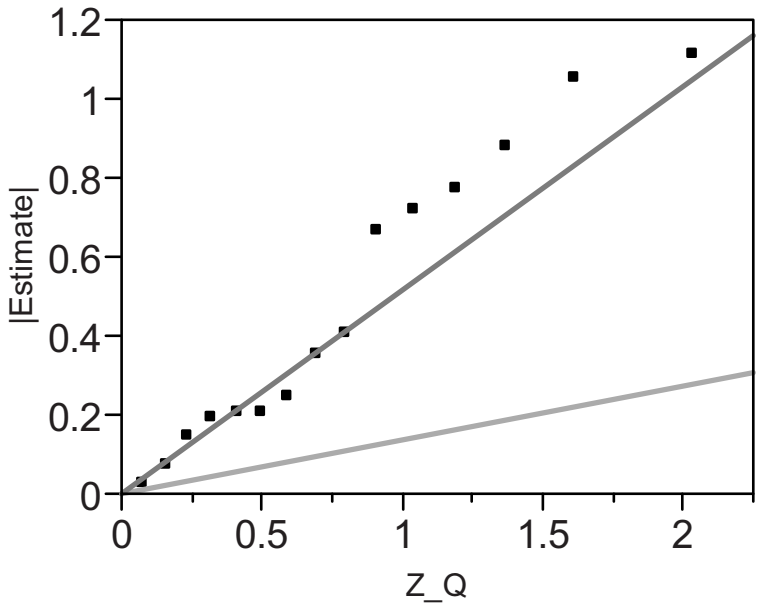


Fig. 2.15. Half-normal plot of effects for Hsieh and Goodwin (1986) data with $y = (\text{pops})^{1/2}$

2.8.3 Modeling variances

In Section 2.6.1, we discussed plotting residuals versus $\hat{y}$ to verify that the assumption of constant variance is reasonable. For non-negative response variables where the ratio $\max(y_i)/\min(y_i)$ is large, we often find it necessary to use a transformation to satisfy the constant variance assumption. In that earlier discussion, however, the mean was primary; checking for equality of variance was a secondary concern. We now consider a different context, in which modeling variability is of primary interest.

Many process improvement applications involve sampling multiple items within each run in order to determine whether within-run variability is smaller at certain treatment combinations. When looking for differences in variability, taking only one or two observations per treatment combination renders an experiment useless. Instead, with primary interest on within-run variability, samples of $m = 10$ or more observations are recommended. For the analysis, one computes the standard deviation s_i or variance s_i^2 for each sample and then proceeds to model this measure of dispersion. If the m values from a sample are independent, normally distributed observations with some mean μ_i and variance σ_i^2, then the sample variance s_i^2 is distributed as a multiple of a chi-square random variable; in particular,

$$s_i^2 \sim [\sigma_i^2/(m-1)]\chi_{m-1}^2$$

and $\text{Var}(s_i^2) = 2\sigma_i^4/(m-1)$. Thus, if we fit a regression model for s_i^2 and have any effects for $E(s_i^2)$, then the constant $\text{Var}(\epsilon)$ assumption will not hold. For this reason, the logarithm is the default variance-stabilizing transformation for standard deviations and variances, since

$$\text{Var}[\ln(s_i^2)] = \text{Var}[\ln(\chi_{m-1}^2)]$$

does not depend on σ_i^2. Bartlett and Kendall (1946) is an early reference regarding the logarithm as a variance-stabilizing transformation for sample variances and standard deviations.

It was mentioned previously that a random sample of size $m \geq 10$ is recommended when studying variation. This is because the precision of a sample variance is poor when the degrees of freedom are few. Given the above result for the chi-squared distribution, the coefficient of variation (CV) for a sample variance of m independent normally distributed observations is $[2/(m-1)]^{1/2}$. Thus, for $m = 10$ observations, the CV is 47%; that is, the standard error for the sample variance is still nearly half as large as the variance we are estimating.

When our response is $\ln(s_i^2)$, the degrees of freedom in s_i^2 determines the variance of the error term in our model. In particular, suppose s_i^2 is the variance of m independent identically distributed observations from a normal distribution; then

$$\text{Var}[\ln(s_i^2)] \approx 2/(m-2).$$

This approximation is excellent for large degrees of freedom and is adequate for m as small as 4. Thus, for a sample of size $m = 10$, we anticipate a RMSE near $[2/(10-2)]^{1/2} = 0.5$ for $\ln(s_i^2)$, or 0.25 for $\ln(s_i)$, provided the data are normally distributed.

Kramschuster et al. (2005) reported two 32-run experiments involving injection molding. For each run, they achieved a steady state and then selected a sample of 10 parts. After aging the parts, $m = 5$ parts per run were carefully measured for shrinkage and warpage. The means of these five observations were effective for finding several active effects for each dimension. However, if one attempts to fit a model to the standard deviations they report, no effects are found. In their case, analysis of the standard deviations is secondary, and measuring five parts per run carefully was quite time-consuming. However, for experiments for which variability is of primary concern, larger samples are generally necessary.

This book does not give any attention to methods for detecting differences in variability from unreplicated designs with no subsampling within runs—even though statisticians have proposed methods for attempting such an analysis. The basic strategy has been to fit a model for the mean, compute residuals, and then use the residuals to discover dispersion effects (i.e., factors that change the variability). For those interested in such methods, see the assessment by Brenneman and Nair (2001). Their concluding remark explains why these methods are not discussed here. "(T)he analysis of location and dispersion effects is intrinsically a difficult problem. In unreplicated experiments, it is really a minefield, one that needs to be maneuvered very carefully. George Box once compared this to trying to squeeze every last bit of water out of a wet towel. If you squeeze too hard, things start breaking down and you can end up making erroneous conclusions" (p. 403).

The first case study in Chapter 4 analyzes a 2^3 factorial with both true replication of runs ($n = 6$) and within-run sampling ($m = 25$), where important differences in within-run variability are found. The samples within each run are unstructured. In some studies of within-run variability, the physical layout suggests likely patterned differences. Section 13.3 discusses advantages of structured samples rather than random samples for variability experiments, and Section 14.3 illustrates the analysis of such data.

2.8.4 Modeling other statistics

Just as count data and variances have default transformations that facilitate the analysis, so do other statistics. Sample correlations r are bounded by the interval $[-1, 1]$, and have more variability when $E(r)$ is near the middle of that range. The default variance-stabilizing transformation for sample correlations, as devised by Sir Ronald Fisher, is

$$f(r) = 0.5 \ln[(1+r)/(1-r)].$$

Recent work by Fujisawa (2000) reinforces this transformation's usefulness.

Ratios (y) constrained to the interval [0, 1] may be transformed using the beta transformation advocated by Rocke (1993):

$$f_B(y;\lambda) = \int_0^y t^{\lambda-1}(1-t)^{\lambda-1}\ \partial t. \tag{2.13}$$

Examples include yield of refining and chemical processes, compositional data, and shrinkage measurements. This beta transformation family includes as special cases the $\arcsin(\sqrt{p})$ and $\ln[p/(1-p)]$ transformations mentioned earlier in Section 2.8.1. Rocke also suggested a generalization of (2.13) where the exponents for t and $1-t$ are allowed to differ.

The logarithm is a useful transformation for lifetime data, t. If the original distribution can be assumed to be lognormal, then $y = \ln(t)$ is normally distributed, and if t follows a two-parameter Weibull distribution, then $\ln(t)$ has an extreme value distribution. In both of these cases, the distribution for $\ln(t)$ is summarized by a location parameter and a scale parameter. Thus, we typically fit a model for $E[\ln(t)]$, with the hope that the variability of the residuals is nearly constant. The interpretability of the fitted model is facilitated by connecting parameters on the $\ln(t)$ scale to parameters of the distribution for t. The log transformation also applies when the response is an order statistic from a lifetime distribution; see, for example, Example 6.5.

Directional response data are often analyzed assuming the von Mises distribution (for responses on a circle) or the von Mises–Fisher distribution (for higher dimensions). Anderson and Wu (1995, 1996) fitted models for both location and dispersion for replicated angular data from a 2^4 factorial design. Anderson-Cook (2001) showed how to model the correlation between an angular response and a continuous response. These methods are relevant for any cyclic response, including time of day (or week or year).

Sometimes the response is a profile or function rather than a scalar. Walker and Wright (2002) analyzed density profiles for fiberboard products. Nair, Taam, and Ye (2002) analyze a compression strength profile for plastic foam. Nair et al. also analyze the audible noise and current of alternators as a function of speed. For each of these examples, the response for the i^{th} experimental run is a sequence of $(y_{ij},\ x_{ij})$, where the x_{ij}'s are univariate and fixed. Assuming the sequence of x_{ij}'s is the same for all runs, one approach is to fit a model for each $j = 1, ..., J$. Nair et al. (2002) took this approach to analyze both the compression strength profiles and the noise output for alternators, in part because no simple functions was adequate to describe the observed data. Shen and Faraway (2004) showed how to conduct inferences for the fitted profiles, whereas Shen and Xu (2007) described diagnostic procedures. A second approach is to fit a curve to the data for each run and then to model some summary measure of each fitted curve. Nair et al. (2002) fitted a three-parameter nonlinear model for each run of the alternator current experiment and then modeled the logarithm of different functions of these parameter estimates. For similar analyses for repeated measures (i.e., longitudinal) data, see Yang, Shen, Xu, and Shoptaw (2007) and Engel (2008).

2.9 Unequal Replication and Unequal Variance

Sometimes a design has unequal replication that was planned. For instance, Snee (1985) replicated 4 of the 16 distinct treatment combinations in an experiment that involved several qualitative factors (which precluded the use of replicated center runs). For such planned imbalance, $\mathbf{X}'\mathbf{X}$ is not diagonal, but it may be block diagonal or have some other structure that may be exploited in the analysis (Dykstra 1959, Liao and Chai 2009).

In other cases, some intended runs fail to produce data, or we discard outlier observations, and end up with unequal replication that is unplanned. Let n_i denote the number of observations at each of the $i = 1, \ldots, 2^k$ treatment combinations. Here we consider the case where $n_i \geq 1$ for all i ($i = 1, \ldots, 2^k$); that is, we have a full factorial with unequal replication. In the next section we consider applications where $n_i = 0$ for some i.

With unequal replication of a full factorial, one can estimate the saturated model (1.4) but due to the lack of balance some regression coefficient estimates change when other terms are dropped from the model. There is some disagreement about which tests are most appropriate (Nelder and Lane 1995, Langsrud 2001). We illustrate the issues using data similar to Dykstra's (1959) 2^3 example. Table 2.9 reports the 12 responses for this experiment that contained replication at half of the treatment combinations.

Table 2.9. Partially replicated 2^3 factorial

x_1	x_2	x_3	Observations
−1	−1	−1	18.4, 20.6
1	−1	−1	25.1
−1	1	−1	24.3
1	1	−1	24.4, 26.2
−1	−1	1	20.4
1	−1	1	25.8, 27.0
−1	1	1	23.6, 24.6
1	1	1	27.9

Fitting a saturated model, we obtain a MSE $= 5.26/4 = 1.315$. The fitted model and t statistics are listed in Table 2.10. The standard error for each coefficient in the saturated model is $\sigma/(10.\bar{6})^{1/2}$, rather than $\sigma/(12)^{1/2}$, due to correlations among pairs of estimates.

Table 2.10. Saturated model for partially replicated 2^3 factorial

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	24.125	0.351	68.71	<.0001
x_1	2.050	0.351	5.84	.0043
x_2	1.275	0.351	3.63	.0221
x_3	0.575	0.351	1.64	.1768
$x_1 * x_2$	−0.850	0.351	−2.42	.0727
$x_1 * x_3$	0.400	0.351	1.14	.3182
$x_2 * x_3$	0.025	0.351	0.07	.9467
$x_1 * x_2 * x_3$	0.300	0.351	0.85	.4410

If the design were balanced, a reduced model could be selected in a single step. One might choose a reduced model with two, three, or four terms depending on whether one retains the $x_1 * x_2$ interaction and whether one follows the practice of retaining all main effects for full factorial designs (as was mentioned in Section 2.4). Regardless of which model is chosen, for balanced designs estimates for the terms in the model are unaffected, as are their t statistics if they are based on the pure error mean square.

Lack of balance complicates the choice of a model. Four possible fitted models are displayed in Figure 2.16. The columns for x_3 and $x_1 * x_2$ are correlated with a correlation of $-1/3$. If both terms are included in the model, as is the case in the reduced model with four terms, the estimates are different than when only one of these terms is included. This causes some ambiguity, since each estimate is larger when the other is omitted. Here, the other columns are orthogonal because of the careful choice of which four runs are replicated. In other nonorthogonal situations, all estimates may be correlated.

So how should one approach model selection? For a full 2^k with unequal replication, stepwise regression procedures are useful. First, fit the saturated model and use backward elimination for models restricted to be hierarchical. Then apply forward selection, again requiring hierarchical models, to see if the same model is obtained. For the data in Table 2.9, using $\alpha = .05$, both procedures lead to reduced model 3 in Figure 2.16.

Unequal replication is particularly common when the responses are from voluntary participants. If the assignment to treatment combinations is made before one knows which participants will respond, then the number of participants contacted needs to be large enough (i) to avoid empty cells and (ii) to avoid large correlations among the columns of the model matrix. Let n denote the number of participants invited per treatment combination (so that, in total, $N = 2^k n$ are invited) and let π denote a (conservative) guess for the proportion of participants who will agree to participate. Then having $n \geq 5/\pi$ is sufficient to avoid empty cells. However, when the realized 2^k sample sizes are random, the distribution for the correlation between two

Parameter Estimates for Reduced Model 1

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	24.0250	0.476	50.46	<.0001
X1	2.0417	0.476	4.29	0.0020
X2	1.1417	0.476	2.40	0.0400

Parameter Estimates for Reduced Model 2

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	24.0250	0.404	59.52	<.0001
X1	2.0417	0.404	5.06	0.0010
X2	1.1417	0.404	2.83	0.0222
X3	0.8583	0.404	2.13	0.0661

Parameter Estimates for Reduced Model 3

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	24.0250	0.346	69.54	<.0001
X1	2.0417	0.346	5.91	0.0004
X2	1.1417	0.346	3.30	0.0108
X1*X2	-1.0417	0.346	-3.01	0.0167

Parameter Estimates for Reduced Model 4

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	24.0250	0.307	78.18	<.0001
X1	2.0417	0.307	6.64	0.0003
X2	1.1417	0.307	3.71	0.0075
X3	0.5750	0.326	1.76	0.1211
X1*X2	-0.8500	0.326	-2.61	0.0350

Fig. 2.16. Four reduced models for partially replicated 2^3

columns of $\mathbf{X}$ is symmetric about 0, with a standard deviation of approximately $[(1-\pi)/(2^k n\pi)]^{1/2}$. For a 2^3, $n\pi = 8$ expected responses per treatment combination may seem sufficient, but if $\pi = 0.1$, $[(1-\pi)/(2^k n\pi)]^{1/2} = 0.12$, so about 5% of the correlations will exceed .24 in magnitude. The smaller the expected proportion π, the larger the expected number responding is required to avoid large correlations.

We now turn to the second topic of this section: unequal variance. If the error variance, σ^2, is not constant, then the least squares estimators will be correlated, even if $\mathbf{X}'\mathbf{X}$ is diagonal. These correlations do not bias the ordinary least squares (OLS) estimators, but they do make OLS inefficient. When the variance of the response is a function of the mean $E(y)$, using a variance-stabilizing transformation resolves this difficulty by changing the model to one where OLS is appropriate. If replication is sufficient to estimate precisely the error variance for each run, then weighted least squares may be applied, weighting by the reciprocal of the estimated variances (see Section 14.4). This standard modification to least squares is discussed in most linear regression books. Its use is unnecessary when the unexplained variation is negligible.

2.10 The Impact of Missing Treatment Combinations

When all 2^k treatment combinations have at least one observation, one can fit the full factorial model (1.4) or any reduced model. In such cases, unequal replication results in correlated estimates of the coefficients but does not alter which models can be fit. Suppose instead that there are $m > 0$ factorial treatment combinations with no data. Then one must omit at least m coefficients from the full factorial model. Because the likelihood of missing treatment combinations is greatest for unreplicated 2^k factorials, we focus on that case.

Our approach to analyzing 2^k factorials with missing observations will be first to fit a saturated hierarchical model. If only one observation is missing, the saturated model is the full factorial model with the k-factor interaction omitted. If two or more observations are missing, there are several options. The details will be shown later.

For this section we use the following notation: $N = 2^k$ is the intended number of runs, of which m are missing, and r is the number of columns for the model matrix $\mathbf{X}$. If $r = N - m$, the model is saturated.

2.10.1 One missing treatment combination

If any single observation is lost from an unreplicated 2^k factorial, $(\mathbf{X}'\mathbf{X})^{-1}$ has a simple structure. Diagonal elements equal $(N - r + 1)/[N(N - r)]$ and off-diagonal elements equal $\pm 1/[N(N-r)]$. For the saturated model ($r = N-1$), this implies that $\text{Var}(b_i) = 2\sigma^2/N$, double what it would have been for the complete 2^k, and all estimates are correlated with a correlation of $\pm.5$. If fewer terms are included in the model, these correlations $\pm 1/(N - r + 1)$ decrease in magnitude and the variances are reduced. Even then, the loss of orthogonality has a much greater impact on the analysis than does the reduction of the sample size.

Draper and Stoneman (1964) present the following example. The full data appear in the Figure 2.17 cube plot, and a half-normal plot of effects is shown in Figure 2.18a. No simple model will account for these data, primarily due to the observation $y = 44$. Upon investigation, it was learned that at the high level for all three factors, "the experimental material changed its form." If this observation is treated as missing in the analysis, the two-factor interaction model can be estimated. Under this saturated model, the predicted value for the (+1, +1, +1) treatment combination is $\hat{y} = 28$, 16 less than 44.

The half-normal plot in Figure 2.18b of the six (correlated) coefficients indicates no effect for x_3. The reduced model

$$\hat{y} = 12.75 + 7.25x_1 - 0.75x_2 + 2.75x_1 * x_2$$

fits the data very well, except near the high level for all factors.

Contrast the two half-normal plots in Figure 2.18. In the second plot, the clump of estimates near zero for the model fitting only the seven treatment

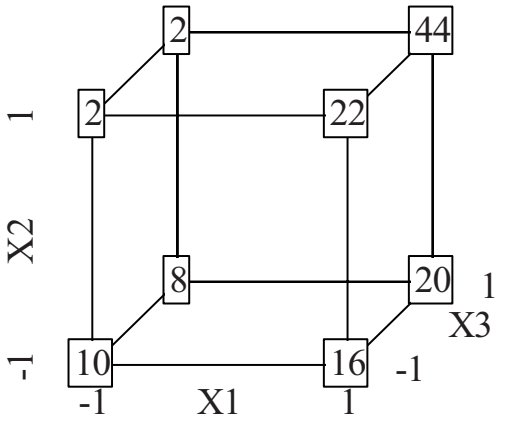


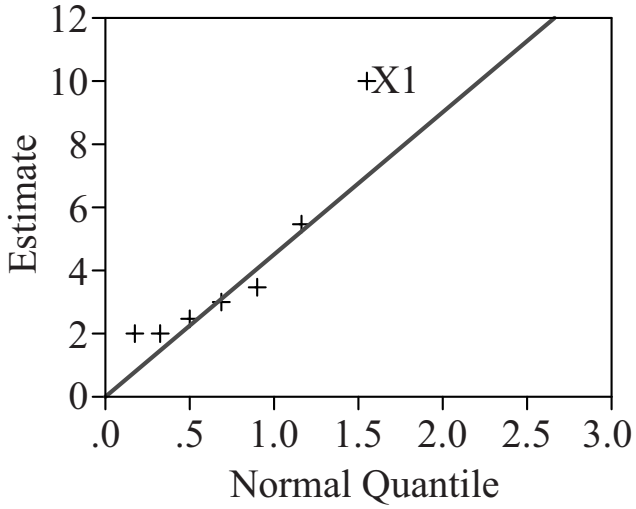
Fig. 2.17. Draper and Stoneman 2^3

combinations indicates that a model with only a few terms will fit very well. In contrast, fitting a model to all eight observations produces a clump of estimates in the range 2–3. Daniel (1959) observed that a single outlier with large error E would affect all the estimated coefficients by $\pm E/N$, pushing a majority of estimates for negligible effects away from zero. Thus, half-normal plots like those in Figures 2.18a and 2.18b are indicative of a simple model accounting for all but one of the observed y values.

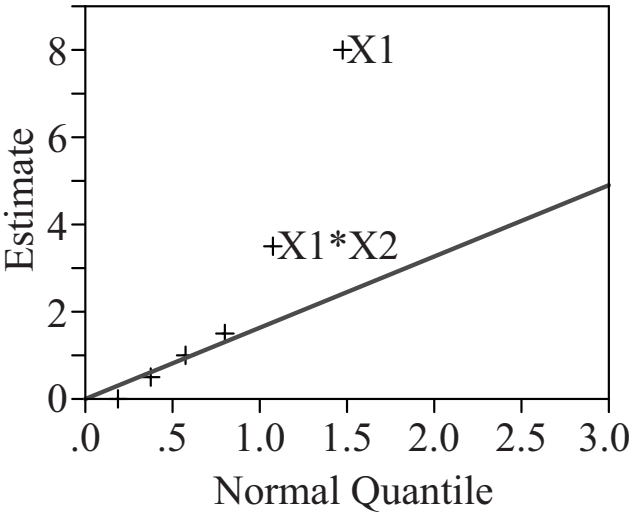
Note that the fitted model corresponding to Figure 18b assumes that the three-factor interaction coefficient is zero. If this assumption were not correct, then all the regression coefficients would be biased by $\pm\beta_{1\cdot 2\cdot 3}$. In general, with one observation missing, the bias for each coefficient from assuming away the highest-order interaction is $\pm\beta_{1\cdot 2\ldots k}/(N-r)$. Thus, a clump of estimates close to zero (as in Figure 2.18b) adds credence to the assumption that the highest order interaction is zero.

2.10.2 Two or more missing treatment combinations

To fit a hierarchical model with $m > 1$ missing observations, there may be several hierarchical saturated models that can be estimated from the data. Using software to fit a model with all terms except the highest-order interaction will result in $m-1$ linear dependencies. Use these "singularity details," as they are labeled in some software, to determine which choices one has for removing $m-1$ additional terms. For each possible saturated model, view the half-normal plot for (correlated) estimated effects. Finding a clump of estimates close to zero is consistent with the assumption that the omitted effects are negligible and that a further simplification of the model is possible. We now illustrate such an analysis for Daniel's drill data (Figure 2.6) by omitting the observed values with the two lowest advance rates (1.68 and 1.98) and the



(a) Half-normal plot for 7 estimates from full factorial



(b) Half-normal plot for 6 correlated estimates

Fig. 2.18. Half-normal plots for Draper and Stoneman estimates

highest advance rate (16.3). Fitting a model with $r = 15$ to the $N - m = 13$ observations produces the following singularities:

$$\begin{aligned}&\text{Intercept} = x_{\text{Flow}} * x_{\text{Speed}} * x_{\text{Mud}} + \cdots \\ &= -x_{\text{Load}} * x_{\text{Flow}} * x_{\text{Speed}} - x_{\text{Load}} * x_{\text{Flow}} * x_{\text{Mud}} - x_{\text{Load}} * x_{\text{Speed}} * x_{\text{Mud}} + \cdots\end{aligned}$$

(which we have simplified by skipping main effects and two-factor interactions). There is no choice regarding the $x_{\text{Flow}} * x_{\text{Speed}} * x_{\text{Mud}}$ interaction; since we have no data at the $(-1, -1, -1)$ combinations for these factors, this term must be omitted. However, because the second singularity involves the other 3 three-factor interactions, this linear dependency may be removed by omitting any one of these interactions. So there are three possible hierarchical models with $r = 13$ that we may estimate (see Table 2.11). Each one results in a model with three significant main effects and b_{Load} never stands out above the clump of estimates near zero.

Table 2.11. Coefficients for three saturated models for $\ln(y)$, treating three observations from Daniel's drill data as "missing"

Term	Model 1	Model 2	Model 3	Std Error
Intercept	1.535	1.535	1.556	$\sigma/4^{1/2}$
Load	0.038	0.039	0.060	$\sigma/8^{1/2}$
Flow	0.307	0.307	0.307	$\sigma/8^{1/2}$
Speed	0.594	0.594	0.594	$\sigma/8^{1/2}$
Mud	0.181	0.181	0.181	$\sigma/8^{1/2}$
Load*Flow	−0.036	−0.036	−0.036	$\sigma/8^{1/2}$
Load*Speed	−0.014	−0.014	−0.014	$\sigma/8^{1/2}$
Load*Mud	0.014	0.014	0.014	$\sigma/8^{1/2}$
Flow*Speed	−0.088	−0.088	−0.067	$\sigma/4^{1/2}$
Flow*Mud	−0.071	−0.070	−0.049	$\sigma/4^{1/2}$
Speed*Mud	−0.014	−0.014	0.007	$\sigma/4^{1/2}$
Load*Flow*Speed	−0.021	−0.021		$\sigma/8^{1/2}$
Load*Flow*Mud	−0.001		0.021	$\sigma/8^{1/2}$
Load*Speed*Mud		0.001	0.021	$\sigma/8^{1/2}$
Flow*Speed*Mud				
Load*Flow*Speed*Mud				

To test for statistical significance requires a modification to Lenth's procedure, since the estimates are correlated. For details, see Edwards and Mee (2008). The success of finding three significant main effects for this example should not diminish the serious loss of information here. If all 16 observations are available, the standard errors are $\sigma/(16)^{1/2}$. The loss of three observations

doubles the standard error for three estimated two-factor interactions in the saturated model, causing a severe loss of power for detecting these effects.

Because of the correlations that result when we fit a saturated model to a factorial with missing observations, it is important to estimate the coefficients using a reduced model. Here, the fitted reduced model is

$$\widehat{\ln(y)} = 1.583 + 0.279x_{\text{Flow}} + 0.566x_{\text{Speed}} + 0.152x_{\text{Mud}}. \tag{2.14}$$

[Compare with (2.11).] The advantage of the estimates in (2.14) is that their standard errors are $\sigma/(11)^{1/2}$, smaller than the standard errors in Table 2.11. However, this benefit comes at a risk of bias to the estimated coefficients, if in fact omitted terms are active.

Because a saturated model can have highly inflated standard errors when treatment combinations are missing, a further step to model selection is to use some form of forward selection regression, adding interaction terms to a main effects model. For our example, fitting models with only one of the two-factor interactions with the large standard errors in Table 2.11 eliminates the largest correlations and enables one to better assess the presence of these terms. Here, no additional useful terms are found.

A final comment is in order. Because the loss of observations is so detrimental to an unreplicated 2^k factorial, such a design is not recommended unless the experimentation and measurement processes are very dependable. If such a design is run and several observations are lost, one may consider a subsequent set of runs to repair the original design. In such cases, it is advisable to run not only the missing observations but also some duplicate treatment combinations that were satisfactory, to account for a possible shift in the process since the initial 2^k was attempted (see Section 9.6).

http://www.springer.com/978-0-387-89102-6

A Comprehensive Guide to Factorial Two-Level Experimentation

Mee, R.

2009, XXIII, 545 p., Hardcover
ISBN: 978-0-387-89102-6