Chapter 5

Linear model formulas, hypothesis tests and confidence intervals

5.1 Linear model formulas

In R linear models are specified using a *model formula*, which is an expression that contains a tilde (the $\sim$ character). The response is on the left-hand side of the tilde, typically as the name of a variable, e.g. `optden`, but it can also be a function of a variable, e.g. `log(BrainWt)`.

The right-hand side of the formula is composed of *model terms* separated by plus signs. In the formulas below we write the response as `y`, continuous covariates as `x`, `z`, `u`, ... and categorical covariates as `f` and `g`. Note that the categorical covariates are assumed to be stored as factors (which includes ordered factors).

Some of the formulas for typical models are:

Simple linear regression The formula

```
y ~ x
```

denotes the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

In the formula shown above, the intercept term is implicit. If you prefer to make it explicit you can write the formula as

```
y ~ 1 + x
```

Regression through the origin If you do not want the intercept term in the model, you must suppress it using the formula

```
y ~ 0 + x
```

or, alternatively,

```
y ~ x - 1
```

The model specified in this way can be written as

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n$$

Notice that you can remove terms, even the implicit intercept, with a negative sign.

Multiple linear regression Multiple covariates can be listed on the right hand side, as in

```
y ~ 1 + x + z + u
```

corresponding to the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 u_i + \epsilon_i, \quad i = 1, \ldots, n$$

Polynomial regression To include polynomial terms in the model you must protect the circumflex operator by surrounding the term with `I()`, which is the identity operator. It implies that the expression inside is to be taken literally in terms of the arithmetic operators, not as the formula language operators. The model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i, \quad i = 1, \ldots, n$$

is written

```
y ~ x + I(x^2) + I(x^3)
```

Another specification for a polynomial regression model uses the `poly()` function which generates *orthogonal polynomial* terms. The fitted responses will be the same from the model shown above and from

```
y ~ poly(x, 3)
```

but the coefficients will be different because they are defined with respect to the orthogonal polynomials. These have some advantages if you are doing the calculations by hand but in practice you don't expect to be doing so.

One categorical covariate The model described as a one-way analysis of variance for the levels of factor, `f`, corresponds to the formula

```
y ~ f
```

Often we use the function `aov()` instead of `lm()` to fit such models. `aov()` is the same as `lm()` except that it puts an extra tag on the fitted model that designates it as only having categorical covariates. This changes, for example, the `summary()` method, which produces an analysis of variance table instead of a summary of the estimated coefficients. The model that is fit is sometimes written as

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, I \; j = 1, \ldots, n_i$$

although it is not fit in that form.

Two categorical covariates, additive The formula for an additive two-factor analysis of variance model,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \quad i = 1, \ldots, I \; j = 1, \ldots, J \; k = 1, \ldots, n_{ij},$$

is

```
y ~ f + g
```

This produces a two-factor analysis of variance table. In the balanced case the analysis of variance table for this model is equivalent to that for the model

```
y ~ g + f
```

in the sense that, although the rows of the table will be in a different order, they are otherwise the same. For unbalanced data the order of the factors is important. The sums of squares in the table are *sequential* sums of squares corresponding to the contribution of the first factor, given the intercept, then the contribution of the second factor, given the first factor and the intercept, and so on. In particular, *blocking factors*, which represent uncontrolable sources of variability, should be listed before experimental factors.

Two categorical covariates, allowing for interactions If the data include replicate observations (more than one observation at the same combination of covariate values) we can fit and analyze a model with interaction terms with a formula like

```
y ~ f + g + f:g
```

where an expression like `f:g` is a two-factor interaction. Similar expressions are used for higher-order interactions. This model can also be expressed as

```
y ~ f * g
```

In general the asterisk operator, (`*`), generates the main effects plus interactions. A three-factor model with all the main effects, two-factor interactions and the three-factor interaction can be written as

```
y ~ f * g * h
```

Combination of continuous and categorical covariates What is sometimes called an *analysis of covariance* model incorporates both categorical and numeric covariates. If there is only one numeric covariate, `x`, then the model can be described in terms of the lines formed by the fitted values on the `y` versus `x` plot. The most common models are the parallel lines (different intercepts, same slope) generated by

```
y ~ f + x
```

and the model in which slopes and intercepts both vary according to the levels of `f`

```
y ~ f * x
```

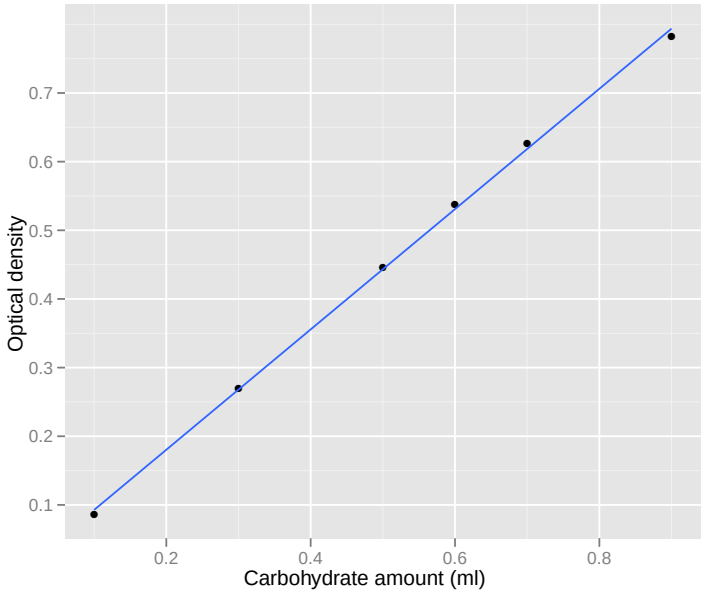

Figure 5.1: Observations of optical density versus carbohydrate amount from the calibration of a Formaldehyde assay.

which is equivalent to

```
y ~ f + x + f:x
```

Occasionally we incorporate an interaction term without a main-effect for `f`.

```
y ~ x + f:x
```

I call this the "zero-dose" model because it is used in the case that `x` represents something like a dose and the levels of `f` corresponds to different treatments. We don't have a main effect for `f` in such a model because a zero dose of treatment 1 is the same as a zero dose of treatment 2. Thus the lines for the different levels of the factors should coincide at `x=0`.

5.2 Examples

The `datasets` package contains several sample datasets that have been used in different texts. By default, this package is attached in an R session.

Simple linear regression The `Formaldehyde` data are a simple example from a calibration study consisting of 6 observations of the carbohydrate content (ml.) (variable `carb`) and the corresponding optical density (variable `optden`). Figure 5.1 is a data plot with the fitted simple linear regression line. This model is fit as

```
> summary(lm1 <- lm(optden ~ 1 + carb, Formaldehyde))

Call:
lm(formula = optden ~ 1 + carb, data = Formaldehyde)
Residuals:
        1         2         3         4         5         6
-0.006714  0.001029  0.002771  0.007143  0.007514 -0.011743

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.005086   0.007834   0.649    0.552
carb        0.876286   0.013535  64.744 3.41e-07

Residual standard error: 0.008649 on 4 degrees of freedom
Multiple R-squared: 0.999,        Adjusted R-squared: 0.9988
F-statistic:  4192 on 1 and 4 DF,  p-value: 3.409e-07
```

(In what follows we will often skip the full summary output and concentrate on the coefficients table, produced by `coef(summary())`, or the analysis of variance table, produced by `anova()`.)

Regression through the origin To constrain the line to pass through the origin (that is, to suppress the `(Intercept)` term) we fit the model as

```
> coef(summary(lm1a <- lm(optden ~ 0 + carb, Formaldehyde)))

      Estimate  Std. Error  t value     Pr(>|t|)
carb 0.8841294 0.005736558 154.1219 2.181654e-10
```

A comparative analysis of variance of these two models

```
> anova(lm1a,lm1)

Analysis of Variance Table
Model 1: optden ~ 0 + carb
Model 2: optden ~ 1 + carb
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1      5 0.00033073
2      4 0.00029920  1 3.1526e-05 0.4215 0.5516
```

produces the same p-value as the t-test on the intercept coefficient in model `lm1`, which is as it should be, because these are two versions of the same test.

Polynomial regression Alternatively, we could fit `optden` as a quadratic function of `carb` using

```
> coef(summary(lm1b <- lm(optden ~ 1 + carb + I(carb^2), Formaldehyde)))
```

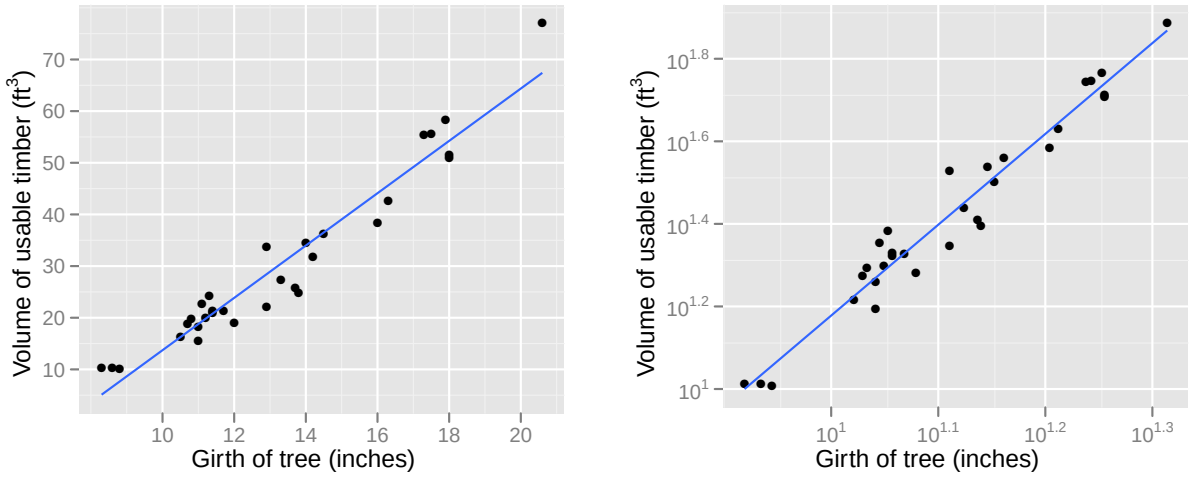

Figure 5.2: Scatterplot of the volume of usable lumber versus the girth of the tree for 31 black cherry trees. The left panel is on the original scale. The right panel is on the log-log scale.

```
              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.0116827  0.0060015 -1.9466   0.14675
carb         0.9711234  0.0267825 36.2596 4.613e-05
I(carb^2)   -0.0962121  0.0263094 -3.6569   0.03532
```

Notice that the quadratic term is significant at the 5% level and generally we would retain it in the model. The reason we don't see much curvature in the data plot (Fig. 5.1) is because there is such a strong linear trend that it masks any nonlinear behaviour.

An alternative specification is

```
> coef(summary(lm1c <- lm(optden ~ poly(carb, 2), Formaldehyde)))

                 Estimate Std. Error  t value  Pr(>|t|)
(Intercept)     0.4578333  0.0017452 262.3429 1.221e-07
poly(carb, 2)1  0.5599550  0.0042748 130.9904 9.810e-07
poly(carb, 2)2 -0.0156326  0.0042748  -3.6569   0.03532
```

Multiple linear regression The `trees` data are measurements of the volume of usable lumber (variable `Volume`) from a sample of 31 black cherry trees. Covariates are a measurement of the girth (`Girth`), which is comparatively easy to measure (you just walk up to the tree and loop a tape measure around it), and the height (`Height`), which is somewhat more difficult to measure. (There is some confusion in the description of the data regarding whether the girth has been converted to an equivalent diameter - we'll assume it is the girth.) If we consider the tree to have the shape of as a cylinder or a cone we would expect that the volume would be related to the square of the girth times the height. In Fig. 5.2 we show the volume versus the girth on the original scale and on a

log-log scale. There is not a tremendous difference in the patterns but careful examination shows better linear behavior in the log-log scale.

Our initial model is

```
> coef(summary(lm2 <- lm(log(Volume) ~ log(Girth), trees)))

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -2.353325   0.230663 -10.202  4.18e-11
log(Girth)   2.199970   0.089835  24.489 < 2.2e-16
```

To fit a model corresponding to a conical or cylindrical shape we add a term in `log(Height)` (recall that we are on the log-log scale)

```
> coef(summary(lm2a <- lm(log(Volume) ~ log(Girth) + log(Height), trees)))

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -6.631617   0.799790 -8.2917 5.057e-09
log(Girth)   1.982650   0.075011 26.4316 < 2.2e-16
log(Height)  1.117123   0.204437  5.4644 7.805e-06
```

Testing specific combinations of parameters At this point we may want to check if a version of the formula for the volume of a cylinder or of a cone, both of which have the form

$$V = k\, d^2 h$$

where k is a constant, d is the diameter (or, equivalently, the girth or circumference at the base) and h is the height. Such an expression would correspond to a value of 2 for the `log(Girth)` coefficient and 1 for the `log(Height)` term. The `log(Height)` term is highly significant and the coefficients of `log(Girth)` and `log(Height)` are reasonably close to 2 and 1. In particular, confidence intervals on these coefficients include 2 and 1

```
> confint(lm2a)

                2.5 %    97.5 %
(Intercept) -8.269912 -4.993322
log(Girth)   1.828998  2.136302
log(Height)  0.698353  1.535894
```

The confidence intervals do not, by themselves, answer the question of whether a model of the form

$$\log(\text{Volume}_i) = \beta_0 + 2\log(\text{Girth}_i) + \log(\text{Height}_i) + \epsilon_i, \quad i = 1, \ldots, 31$$

is a reasonable fit. To fit this model we use an `offset` expression in the model formula.

```
> lm2c <- lm(log(Volume) ~ 1 + offset(2*log(Girth) + log(Height)), trees)
```

and perform a comparative analysis of variance

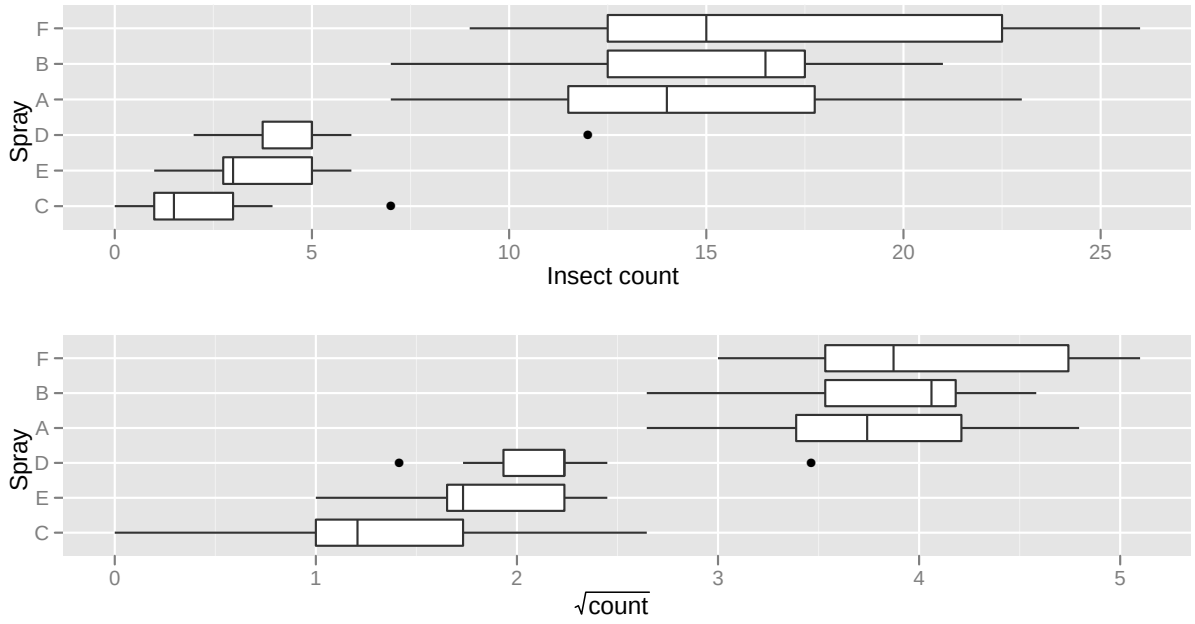

Figure 5.3: Comparative boxplots of the insect count by spray type in the `InsectSprays` data. The sprays have been reordered according to increasing mean response. In the lower panel the response is the square root of the count.

```
> anova(lm2c,lm2a)

Analysis of Variance Table
Model 1: log(Volume) ~ 1 + offset(2 * log(Girth) + log(Height))
Model 2: log(Volume) ~ log(Girth) + log(Height)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     30 0.18769
2     28 0.18546  2 0.0022224 0.1678 0.8464
```

The large p-value indicates that the more complex model (general values of the coefficients for `log(Girth)` and `log(Height)`) does not fit significantly better than the simpler model (assuming `2*log(Girth) + log(Height)`), thus we prefer the simpler model.

One-way analysis of variance Next consider the `InsectSprays` data with a response, `count`, related to a categorical covariate, `spray`. Comparative boxplots (Fig. 5.3) show that the square root of the count is a more reasonable scale for the response and that there is considerable differences in the response according to the spray type.

Althought we can fit a model with categorical covariates using the `lm()` function, there is an advantage in using the `aov()` function instead, because it allows us to extract some additional infor-

mation that applies only to categorical factors. Also, `summary()` applied to an `aov()` model produces the analysis of variance table, which for such models, is more interesting than the coefficients table.

```
> summary(av1 <- aov(sqrt(count) ~ spray, InsectSprays))

            Df Sum Sq Mean Sq F value    Pr(>F)
spray        5 88.438 17.6876  44.799 < 2.2e-16
Residuals   66 26.058  0.3948
```

If we want to express the model in terms of "effects" we can obtain these as

```
> model.tables(av1)

Tables of effects
 spray
spray
      A       B       C       D       E       F
 0.9482  1.0642 -1.5676 -0.6481 -1.0030  1.2062
```

or, if we are interested in the estimates of the means for each group,

```
> model.tables(av1, type="means")

Tables of means
Grand mean
2.812433

 spray
spray
    A     B     C     D     E     F
3.761 3.877 1.245 2.164 1.809 4.019
```

Various types of "multiple comparisons" methods are also available. We will discuss these later.

Multi-factor analysis of variance When we have more than one categorical covariate, as in the `OrchardSprays` data,

```
> str(OrchardSprays)

'data.frame':        64 obs. of  4 variables:
 $ decrease : num  57 95 8 69 92 90 15 2 84 6 ...
 $ rowpos   : num  1 2 3 4 5 6 7 8 1 2 ...
 $ colpos   : num  1 1 1 1 1 1 1 1 2 2 ...
 $ treatment: Factor w/ 8 levels "A","B","C","D",..: 4 5 2 8 7 6 3 1 3 2 ...
```

we simply include them in the model formula. It happens that for this experiment there are two blocking factors, `rowpos` and `colpos`, and one experimental factor, `treatment`, so we put the blocking factors first.

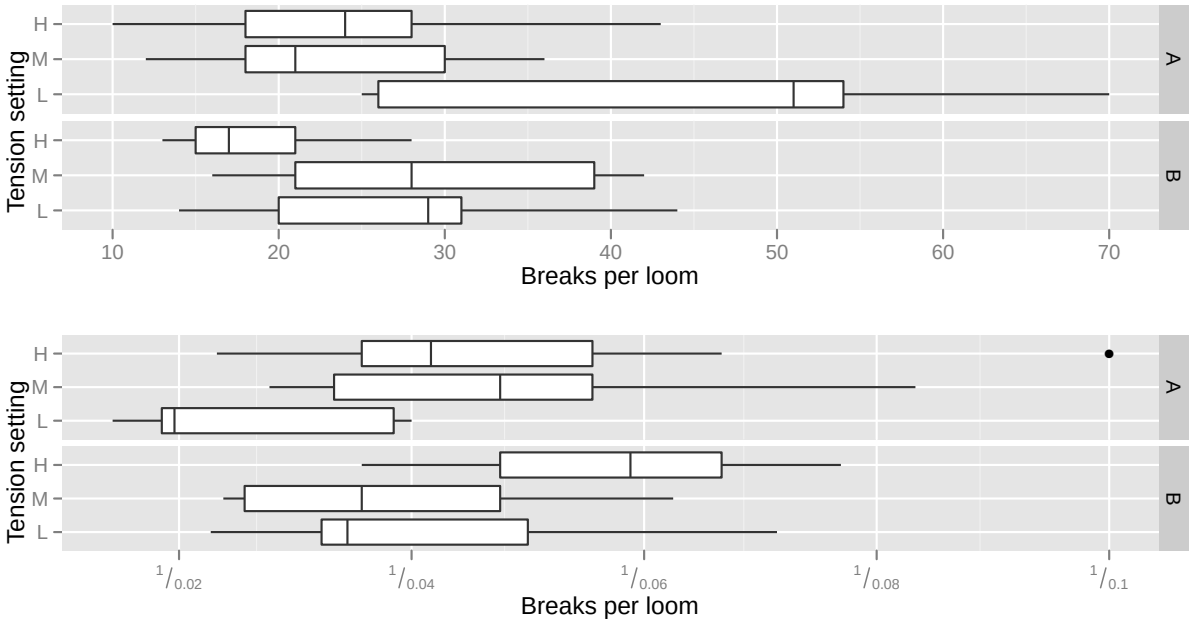

Figure 5.4: Comparative boxplots of the number of warp breaks per loom by tension setting for the `warpbreaks` data. Panels are determined by wool type. The upper two panels are on the original scale of number of breaks. The lower two panels are on the reciprocal scale (i.e. number of looms per break).

```
> summary(av2 <- aov(decrease ~ factor(rowpos) + factor(colpos) + treatment, OrchardSprays))

               Df Sum Sq Mean Sq F value    Pr(>F)
factor(rowpos)  7   4767   681.1  1.7884    0.1151
factor(colpos)  7   2807   401.0  1.0530    0.4100
treatment       7  56160  8022.9 21.0667 7.455e-12
Residuals      42  15995   380.8
```

These data are arranged in what is called a "Latin square" design, which is a special type of fractional replication. There are 64 observations on three factors, each at 8 levels, so not only are there no replications, we don't even have an observation in each of the possible $8 \times 8 \times 8 = 512$ combinations, and cannot try to fit a model with interaction terms.

Multi-factor anova with replications The `warpbreaks` data, shown in Fig. 5.4, are counts of the number of warp breaks per loom (a length of wool) according to the tension setting for the wool and the type of wool. We see that on the original scale of the number of breaks per loom there is increasing variance with an increasing level of the response, whereas on the reciprocal scale (number of looms per break) the variability is much closer to being constant.

Because there are 9 replications at each of the wool/tension combinations

```
> xtabs(~ wool + tension, warpbreaks)

    tension
wool L M H
   A 9 9 9
   B 9 9 9
```

we can fit a model with main effects for `wool` and for `tension` and the `wool:tension` interaction.

```
> summary(av3 <- aov(breaks ~ wool * tension, warpbreaks))

             Df Sum Sq Mean Sq F value    Pr(>F)
wool          1  450.7  450.67  3.7653 0.0582130
tension       2 2034.3 1017.13  8.4980 0.0006926
wool:tension  2 1002.8  501.39  4.1891 0.0210442
Residuals    48 5745.1  119.69
```

In this model the interaction is significant. When an interaction is significant we typically retain both of the main effects in the model.

However, if we fit the model on the reciprocal scale

```
> summary(av3a <- aov(1/breaks ~ wool * tension, warpbreaks))

             Df    Sum Sq    Mean Sq F value   Pr(>F)
wool          1 0.0002403 0.00024035  0.9001 0.347511
tension       2 0.0033455 0.00167274  6.2642 0.003826
wool:tension  2 0.0012088 0.00060442  2.2635 0.114978
Residuals    48 0.0128174 0.00026703
```

we no longer have a significant interaction and could reduce the model to the main effects only

```
> summary(av3b <- aov(1/breaks ~ tension +  wool, warpbreaks))

            Df    Sum Sq    Mean Sq F value   Pr(>F)
tension      2 0.0033455 0.00167274  5.9629 0.004758
wool         1 0.0002403 0.00024035  0.8568 0.359087
Residuals   50 0.0140262 0.00028052
```

Here we have reordered the factors `tension` and `wool` so that `wool` is the last term and thus the second row of the analysis of variance table corresponds to a test of the main effect of the wool given that tension had been taken into account. (If you look closely at the sums of squares, degrees of freedom and mean squares you will see that they are consistent in models `av3b` and `av3a` but that is a consequence of the data being completely balanced with respect to these factors. To be safe, always make the factor you are going to test be the last one in the model formula.) The `wool` factor is not significant and we can reduce the model to a single factor model

```
> summary(av3c <- aov(1/breaks ~ tension, warpbreaks))
```

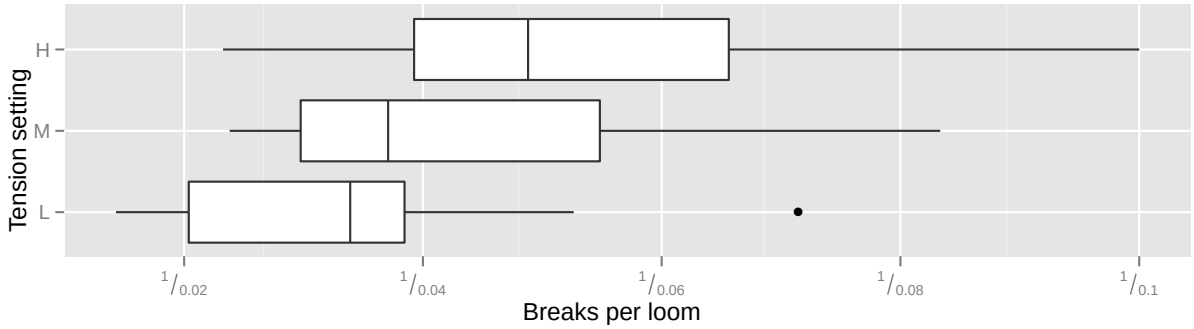

Figure 5.5: Comparative boxplots of the number of warp breaks per loom by tension setting for the `warpbreaks` data. Panels are determined by wool type. The upper two panels are on the original scale of number of breaks. The lower two panels are on the reciprocal scale (i.e. number of looms per break).

```
            Df    Sum Sq    Mean Sq F value   Pr(>F)
tension      2 0.0033455 0.00167274  5.9797 0.004645
Residuals   51 0.0142666 0.00027974
```

corresponding to Fig. 5.5 in which we can see a trend across the three ordered levels of tension; low tension gives a low reciprocal number of breaks (corresponding to a higher frequency of breaks), medium tension gives an intermediate reciprocal number and high tension gives the highest reciprocal number.

This is a common situation with a factor like `tension` whose levels are in a natural ordering, `L < M < H`. Details will be given later but, for now, it is enough to see that if we convert the factor to an ordered factor

```
> str(warpbreaks <- within(warpbreaks, tension <- ordered(tension)))

'data.frame':        54 obs. of  3 variables:
 $ breaks : num  26 30 54 25 70 52 51 26 67 18 ...
 $ wool   : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
 $ tension: Ord.factor w/ 3 levels "L"<"M"<"H": 1 1 1 1 1 1 1 1 1 2 ...
```

and fit the model as before,

```
> summary(av3d <- aov(1/breaks ~ tension, warpbreaks))

            Df    Sum Sq    Mean Sq F value   Pr(>F)
tension      2 0.0033455 0.00167274  5.9797 0.004645
Residuals   51 0.0142666 0.00027974
```

we get the same analysis of variance table but now the two degrees of freedom for `tension` are divided into a linear trend and a quadratic relationship in addition to the linear trend

```
> coef(summary.lm(av3d))

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  4.2798e-02  2.2760e-03 18.8040 < 2.2e-16
tension.L    1.3633e-02  3.9422e-03  3.4582  0.001106
tension.Q   -8.0425e-05  3.9422e-03 -0.0204  0.983803
```

With a p-value of 98.4%, the quadratic term is not at all significant, indicating that we could reduce to only the linear trend.

5.3 Classes and methods for linear models

A model fit with `lm()` has class `"lm"`

```
> class(lm2)

[1] "lm"
```

for which there are several methods defined

```
> methods(class="lm")

 [1] add1.lm*           alias.lm*          anova.lm           case.names.lm*
 [5] confint.lm*        cooks.distance.lm* deviance.lm*       dfbeta.lm*
 [9] dfbetas.lm*        drop1.lm*          dummy.coef.lm*     effects.lm*
[13] extractAIC.lm*     family.lm*         formula.lm*        fortify.lm
[17] hatvalues.lm       influence.lm*      kappa.lm           labels.lm*
[21] logLik.lm*         model.frame.lm     model.matrix.lm    plot.lm
[25] predict.lm         print.lm           proj.lm*           residuals.lm
[29] rstandard.lm       rstudent.lm        simulate.lm*       summary.lm
[33] variable.names.lm* vcov.lm*
   Non-visible functions are asterisked
```

We have already seen several of these in use:

anova Return the (sequential) analysis of variance table for a single fitted model or a comparative analysis of variance for multiple fitted models.

confint Return confidence intervals on the coefficients

deviance Return the residual sum of squares (RSS) for the model. (This is a misnomer because the RSS is related to but not exactly the same as the deviance.)

formula Return the model formula.

kappa Return the condition number of the model matrix or an upper bound on its condition number.

logLik Return the value of the log-likelihood at the estimated parameter values.

model.frame Return the model frame to which the model was actually fit.

model.matrix Return the model matrix.

plot Produce some common residual plots for evaluating the model fit.

predict Returns evaluations of the fitted model, and optionally their standard errors, at the observed or newly specified values of the covariates.

residuals Returns the residuals from the fit.

rstandard Returns the "standardized residuals" (to be described later).

rstandard Returns the "Studentized residuals" (to be described later).

simulate Return a matrix of simulated response vectors according to the model assuming that the fitted values of the parameters are the true parameter values.

summary Return a summary of the fitted model

vcov Return the (estimated) variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ (i.e. the matrix that could be expressed as $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$).

Other extractor functions such as `coef` and `fitted` do not have specific methods for class `"lm"` but instead apply the default method to objects of this class.

A model fit by `aov` has class

```
> class(av3)

[1] "aov" "lm"
```

`"aov"` and also class `"lm"`. This means that methods for class `"aov"` will be chosen, if they exist, otherwise methods for class `"lm"` and, finally, the default method.

Specific methods for class `"aov"` are

```
> methods(class="aov")

[1] coef.aov*         extractAIC.aov*   model.tables.aov* print.aov*
[5] proj.aov*         se.contrast.aov*  summary.aov       TukeyHSD.aov
   Non-visible functions are asterisked
```

from which we can see that specific methods for the `coef`, `extractAIC`, `print`, `proj` and `summary` generics are available for this class and will be chosen in preference to the `"lm"` or default method. Furthermore there are specific methods for the `model.tables`, `se.contrast` and `TukeyHSD` generics.

We have seen the use of `model.tables` and will later use `TukeyHSD` which returns adjusted confidence intervals on differences between levels using Tukey's "Honest Significant Difference" technique. This is one of the multiple comparisons techniques mentioned earlier.

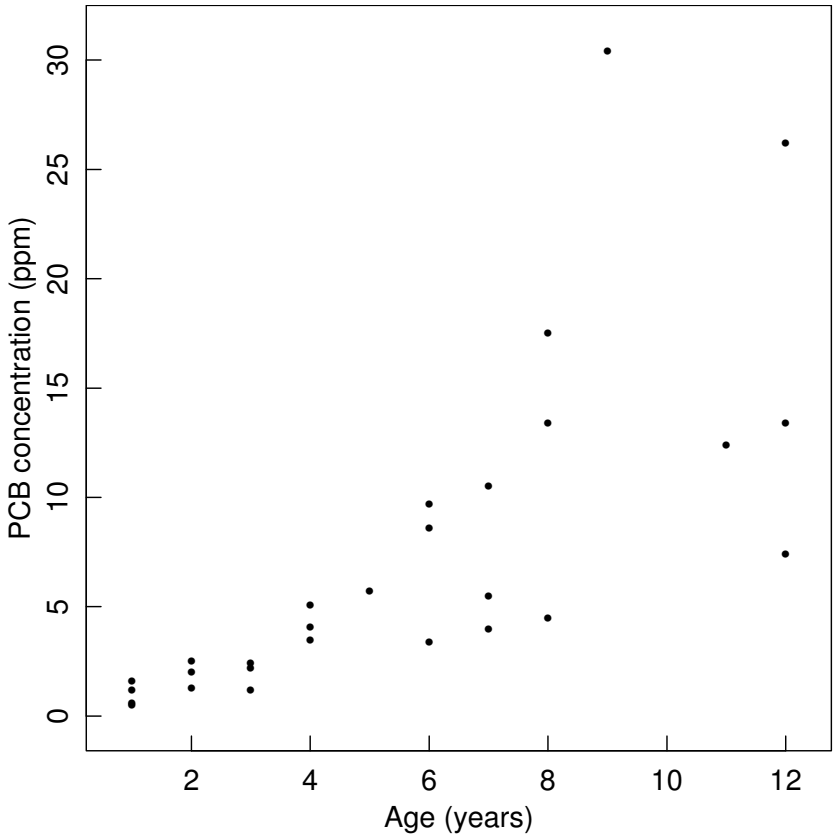

Figure 5.6: Plot of PCB concentration versus age for lake trout.

5.4 Geometry of least squares estimators

Example 1. *As a simple example of a linear regression model, we consider the concentration of polychlorinated biphenyls (PCBs) in Lake Cayuga trout as a function of age Bache, Serum, Youngs, and Lisk (1972). The data set is described in Appendix 1, Section A1.1. A plot of the PCB concentration versus age, Figure 5.6, reveals a curved relationship between PCB concentration and age. Furthermore, there is increasing variance in the PCB concentration as the concentration increases. Since the assumption (**??**) requires that the variance of the disturbances be constant, we seek a transformation of the PCB concentration which will stabilize the variance (see Section 1.3.2). Plotting the PCB concentration on a logarithmic scale, as in Figure 5.7a, nicely stabilizes the variance and produces a more nearly linear relationship. Thus, a linear expectation function of the form*

$$\ln(\text{PCB}) = \beta_1 + \beta_2 \,\text{age}$$

could be considered appropriate, where ln *denotes the natural logarithm (logarithm to the base e). Transforming the regressor variable Box and Tidwell (1962) can produce an even straighter plot, as shown in Figure 5.7b, where we use the cube root of age. Thus a simple expectation function to be*

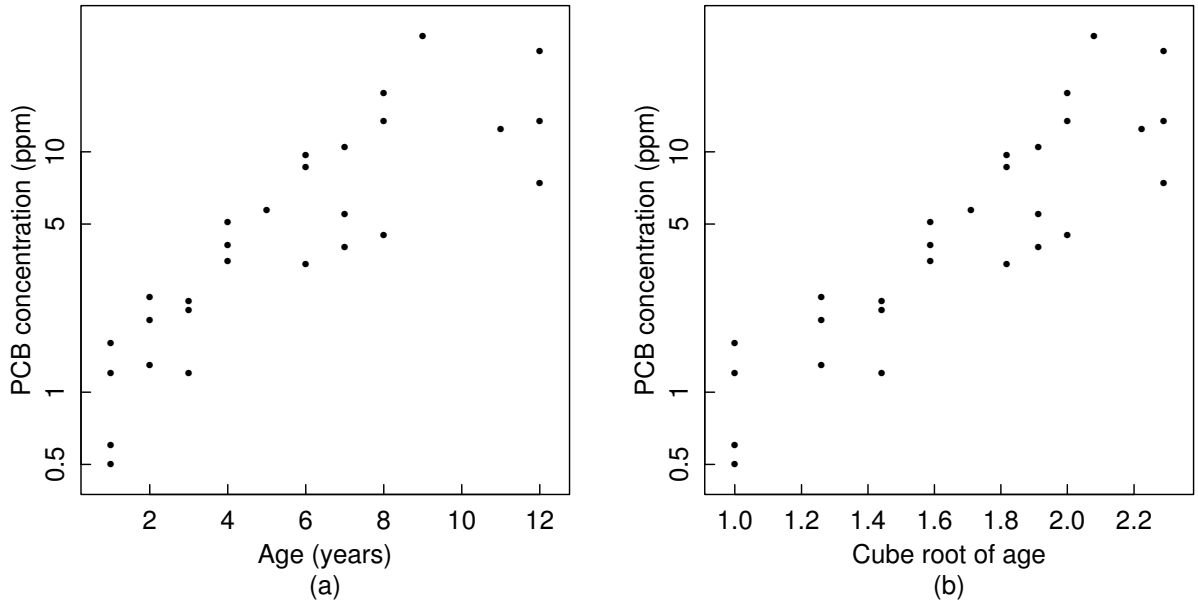

Figure 5.7: Plot of PCB concentration versus age for lake trout. The concentration, on a logarithmic scale, is plotted versus age in part *a* and versus $\sqrt[3]{\text{age}}$ in part *b*.

fitted is

$$\ln(\text{PCB}) = \beta_1 + \beta_2 \sqrt[3]{age}$$

(Note that the methods of Chapter 2 can be used to fit models of the form

$$f(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \beta_0 + \beta_1 x_1^{\alpha_1} + \beta_2 x_2^{\alpha_2} + \cdots + \beta_P x_P^{\alpha_P}$$

by simultaneously estimating the conditionally linear parameters $\boldsymbol{\beta}$ and the transformation parameters $\boldsymbol{\alpha}$. The powers $\alpha_1, \ldots, \alpha_P$ are used to transform the factors so that a simple linear model in $x_1^{\alpha_1}, \ldots, x_P^{\alpha_P}$ is appropriate. In this book we use the power $\alpha = 0.33$ for the age variable even though, for the PCB data, the optimal value is 0.20.)

5.4.1 The Least Squares Estimates

The *likelihood function*, or more simply, the *likelihood*, $l(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$, for $\boldsymbol{\beta}$ and σ is identical in form to the joint probability density (**??**) except that $l(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$ is regarded as a function of the parameters conditional on the observed data, rather than as a function of the responses conditional on the values of the parameters. Suppressing the constant $(2\pi)^{-N/2}$ we write

$$l(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) \propto \sigma^{-N} \exp\left(\frac{-\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \tag{5.1}$$

The likelihood is maximized with respect to $\boldsymbol{\beta}$ when the *residual sum of squares*

$$\begin{aligned} S(\boldsymbol{\beta}) &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \\ &= \sum_{n=1}^{N} \left[y_n - \left(\sum_{p=1}^{P} x_{np}\beta_p \right) \right]^2 \end{aligned} \tag{5.2}$$

is a minimum. Thus the *maximum likelihood estimate* $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ which minimizes $S(\boldsymbol{\beta})$. This $\hat{\boldsymbol{\beta}}$ is called the *least squares* estimate and can be written

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{5.3}$$

Least squares estimates can also be derived by using sampling theory, since the least squares estimator is the minimum variance unbiased estimator for $\boldsymbol{\beta}$, or by using a Bayesian approach with a noninformative prior density on $\boldsymbol{\beta}$ and σ. In the Bayesian approach, $\hat{\boldsymbol{\beta}}$ is the mode of the marginal posterior density function for $\boldsymbol{\beta}$.

All three of these methods of inference, the likelihood approach, the sampling theory approach, and the Bayesian approach, produce the same point estimates for $\boldsymbol{\beta}$. As we will see shortly, they also produce similar regions of "reasonable" parameter values. First, however, it is important to realize that the least squares estimates are only appropriate when the model (**??**) and the assumptions on the disturbance term, (**??**) and (**??**), are valid. Expressed in another way, in using the least squares estimates we assume:

1. The expectation function is correct.

2. The response is expectation function plus disturbance.

3. The disturbance is independent of the expectation function.

4. Each disturbance has a normal distribution.

5. Each disturbance has zero mean.

6. The disturbances have equal variances.

7. The disturbances are independently distributed.

When these assumptions appear reasonable and have been checked using diagnostic plots such as those described in Section 1.3.2, we can go on to make further inferences about the regression model.

Looking in detail at each of the three methods of statistical inference, we can characterize some of the properties of the least squares estimates.

5.4.2 Sampling Theory Inference Results

The least squares estimator has a number of desirable properties as shown, for example, in Seber (1977):

1. The least squares estimator $\hat{\boldsymbol{\beta}}$ is normally distributed. This follows because the estimator is a linear function of $\boldsymbol{Y}$, which in turn is a linear function of $\boldsymbol{Z}$. Since $\boldsymbol{Z}$ is assumed to be normally distributed, $\hat{\boldsymbol{\beta}}$ is normally distributed.
2. $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$: the least squares estimator is unbiased.
3. $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$: the covariance matrix of the least squares estimator depends on the variance of the disturbances and on the derivative matrix $\boldsymbol{X}$.
4. A $1-\alpha$ *joint confidence region* for $\boldsymbol{\beta}$ is the ellipsoid
$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq Ps^2F_{P,N-P;\alpha} \tag{5.4}$$
where
$$s^2 = \frac{S(\widehat{\boldsymbol{\beta}})}{N-P}$$
is the *residual mean square* or *variance estimate* based on $N-P$ *degrees of freedom*, and $F_{P,N-P;\alpha}$ is the upper α quantile for Fisher's F distribution with P and $N-P$ degrees of freedom.
5. A $1-\alpha$ *marginal confidence interval* for the parameter β_p is
$$\hat{\beta}_p \pm \mathrm{se}(\hat{\beta}_p)t_{N-P;\alpha} \tag{5.5}$$
where $t_{N-P;\alpha}$ is the upper $\alpha/2$ quantile for Student's T distribution with $N-P$ degrees of freedom and the standard error of the parameter estimator is
$$\mathrm{se}(\hat{\beta}_p) = s\sqrt{\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\}_{pp}} \tag{5.6}$$
with $\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}_{pp}$ equal to the pth diagonal term of the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
6. A $1-\alpha$ confidence interval for the expected response at $\boldsymbol{x}_0$ is
$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\; t_{N-P;\alpha} \tag{5.7}$$
7. A $1-\alpha$ confidence interval for the expected response at $\boldsymbol{x}_0$ is
$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\; t_{N-P;\alpha} \tag{5.8}$$
8. A $1-\alpha$ confidence band for the response function at any $\boldsymbol{x}$ is given by
$$\boldsymbol{x}'\hat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}\sqrt{PF_{P,N-P;\alpha}} \tag{5.9}$$

The expressions (5.8) and (5.9) differ because (5.8) concerns an interval at a single specific point, whereas (5.9) concerns the band produced by the intervals at all the values of $\boldsymbol{x}$ considered simultaneously.

5.4.3 Likelihood Inference Results

The likelihood $l(\boldsymbol{\beta}, \sigma \mid \boldsymbol{y})$, equation (5.1), depends on $\boldsymbol{\beta}$ only through $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|$, so likelihood contours are of the form

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = c \tag{5.10}$$

where c is a constant. A likelihood region bounded by the contour for which

$$c = \ S(\hat{\boldsymbol{\beta}}) \left[1 + \frac{P}{N-P} F_{P,N-P;\alpha}\right]$$

is identical to a $1-\alpha$ joint confidence region from the sampling theory approach. The interpretation of a likelihood region is quite different from that of a confidence region, however.

5.4.4 Bayesian Inference Results

As shown in Box and Tiao (1973), the Bayesian marginal posterior density for $\boldsymbol{\beta}$, assuming a noninformative prior density for $\boldsymbol{\beta}$ and σ of the form

$$p(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1} \tag{5.11}$$

is

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \left\{1 + \ (\boldsymbol{\beta} \ - \ \hat{\boldsymbol{\beta}})' \boldsymbol{X}' \boldsymbol{X} \frac{(\boldsymbol{\beta} - \ \hat{\boldsymbol{\beta}})}{\nu s^2}\right\}^{-(\nu+P)/2} \tag{5.12}$$

which is in the form of a P-variate Student's T density with *location parameter* $\hat{\boldsymbol{\beta}}$, *scaling matrix* $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and $\nu = N - P$ degrees of freedom. Furthermore, the marginal posterior density for a single parameter β_p, say, is a univariate Student's T density with location parameter $\hat{\beta}_p$, scale parameter $s^2 \left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}_{pp}$, and degrees of freedom $N - P$. The marginal posterior density for the mean of y at $\boldsymbol{x}_0$ is a univariate Student's T density with location parameter $\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$, scale parameter $s^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$, and degrees of freedom $N - P$.

A *highest posterior density* (HPD) region of content $1 - \alpha$ is defined Box and Tiao (1973) as a region R in the parameter space such that $\Pr\{\boldsymbol{\beta} \in \mathrm{R}\} = 1 - \alpha$ and, for $\boldsymbol{\beta}_1 \in \mathrm{R}$ and $\boldsymbol{\beta}_2 \notin \mathrm{R}$, $p(\boldsymbol{\beta}_1|\boldsymbol{y}) \geq p(\boldsymbol{\beta}_2|\boldsymbol{y})$. For linear models with a noninformative prior, an HPD region is therefore given by the ellipsoid defined in (1.9). Similarly, the marginal HPD regions for β_p and $\boldsymbol{x}_0'\boldsymbol{\beta}$ are numerically identical to the sampling theory regions (5.6, 5.7, and 5.8).

5.4.5 Comments

Although the three approaches to statistical inference differ considerably, they lead to essentially identical inferences. In particular, since the joint confidence, likelihood, and Bayesian HPD regions are identical, we refer to them all as *inference regions*.

In addition, when referring to standard errors or correlations, we will use the Bayesian term "the standard error of β_p" when, for the sampling theory or likelihood methods, we should more properly say "the standard error of the estimate of β_p".

For linear least squares, any of the approaches can be used. For nonlinear least squares, however, the likelihood approach has the simplest and most direct geometrical interpretation, and so we emphasize it.

Example 2. *The PCB data can be used to determine parameter estimates and joint and marginal inference regions. In this linear situation, the regions can be summarized using* $\hat{\boldsymbol{\beta}}$, s^2, $\boldsymbol{X}'\boldsymbol{X}$, *and* $\nu = N - P$. *For the* ln(PCB) *data with* $\sqrt[3]{age}$ *as the regressor, we have* $\hat{\boldsymbol{\beta}} = (-2.391, 2.300)'$, $s^2 = 0.246$ *on* $\nu = 26$ *degrees of freedom, and*

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{X} &= \begin{bmatrix} 28.000 & 46.941 \\ 46.941 & 83.367 \end{bmatrix} \\ (\boldsymbol{X}'\boldsymbol{X})^{-1} &= \begin{bmatrix} 0.6374 & -0.3589 \\ -0.3589 & 0.2141 \end{bmatrix} \end{aligned}$$

The joint 95% inference region is then

$$\begin{aligned} 28.00(\beta_1 + 2.391)^2 + 93.88(\beta_1 + 2.391)(\beta_2 - 2.300) + 83.37(\beta_2 - 2.300)^2 &= 2(0.246)3.37 \\ &= 1.66 \end{aligned}$$

the marginal 95% inference interval for the parameter β_1 *is*

$$-2.391 \pm (0.496)\sqrt{0.6374}(2.056)$$

or

$$-3.21 \leq \beta_1 \leq -1.58$$

and the marginal 95% inference interval for the parameter β_2 *is*

$$2.300 \pm (0.496)\sqrt{0.2141}(2.056)$$

or

$$1.83 \leq \beta_2 \leq 2.77$$

The 95% inference band for the ln(PCB) *value at any* $\sqrt[3]{age} = x$, *is*

$$-2.391 + 2.300x \pm (0.496)\sqrt{0.637 - 0.718x + 0.214x^2}\sqrt{2(3.37)}$$

These regions are plotted in Figure 5.8.

While it is possible to give formal expressions for the least squares estimators and the regression summary quantities in terms of the matrices $\boldsymbol{X}'\boldsymbol{X}$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, the use of these matrices for computing the estimates is not recommended. Superior computing methods are presented in Section 1.2.2.

Finally, the assumptions which lead to the use of the least squares estimates should always be examined when using a regression model. Further discussion on assumptions and their implications is given in Section 1.3.

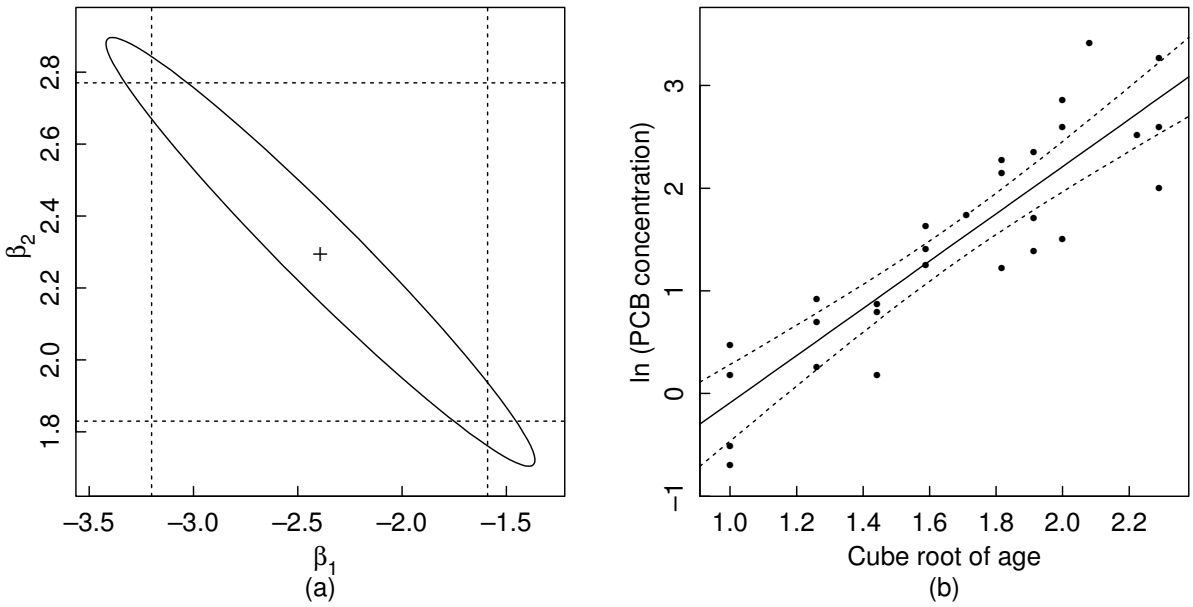

Figure 5.8: Inference regions for the model $\ln(\text{PCB}) = \beta_1 + \beta_2 \sqrt[3]{\text{age}}$. Part *a* shows the least squares estimates (+), the parameter joint 95% inference region (solid line), and the marginal 95% inference intervals (dotted lines). Part *b* shows the fitted response (solid line) and the 95% inference band (dotted lines).

5.5 The Geometry of Linear Least Squares

The model (**??**) and assumptions (**??**) and (**??**) lead to the use of the least squares estimate (5.3) which minimizes the residual sum of squares (5.2). As implied by (5.2), $S(\boldsymbol{\beta})$ can be regarded as the square of the distance from the data vector $\boldsymbol{y}$ to the expected response vector $\boldsymbol{X\beta}$. This links the subject of linear regression to Euclidean geometry and linear algebra. The assumption of a normally distributed disturbance term satisfying (**??**) and (**??**) indicates that the appropriate scale for measuring the distance between $\boldsymbol{y}$ and $\boldsymbol{X\beta}$ is the usual Euclidean distance between vectors. In this way the Euclidean geometry of the N-dimensional response space becomes statistically meaningful. This connection between geometry and statistics is exemplified by the use of the term *spherical normal* for the normal distribution with the assumptions (**??**) and (**??**), because then contours of constant probability are spheres.

Note that when we speak of the linear form of the expectation function $\boldsymbol{X\beta}$, we are regarding it as a function of the parameters $\boldsymbol{\beta}$, and that when determining parameter estimates we are only concerned with how the expected response depends on the *parameters*, not with how it depends on the *variables*. In the PCB example we fit the response to $\sqrt[3]{\text{age}}$ using linear least squares because the parameters $\boldsymbol{\beta}$ enter the model linearly.

5.5.1 The Expectation Surface

The process of calculating $S(\boldsymbol{\beta})$ involves two steps:

1. Using the P-dimensional parameter vector $\boldsymbol{\beta}$ and the $N \times P$ derivative matrix $\boldsymbol{X}$ to obtain the N-dimensional *expected response vector* $\boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{X\beta}$, and

2. Calculating the squared distance from $\boldsymbol{\eta}(\boldsymbol{\beta})$ to the observed response $\boldsymbol{y}$, $\|\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta})\|^2$.

The possible expected response vectors $\boldsymbol{\eta}(\boldsymbol{\beta})$ form a P-dimensional *expectation surface* in the N-dimensional response space. This surface is a linear subspace of the response space, so we call it the *expectation plane* when dealing with a linear model.

Example 3. *To illustrate the geometry of the expectation surface, consider just three cases from the* ln(PCB) *versus* $\sqrt[3]{age}$ *data,*

$\sqrt[3]{age}$	ln(PCB)
1.26	*0.92*
1.82	*2.15*
2.22	*2.52*

The matrix $\boldsymbol{X}$ *is then*

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.26 \\ 1 & 1.82 \\ 1 & 2.22 \end{bmatrix}$$

which consists of two column vectors $\boldsymbol{x}_1 = (1, 1, 1)'$ *and* $\boldsymbol{x}_2 = (1.26, 1.82, 2.22)'$. *These two vectors in the 3-dimensional response space are shown in Figure 5.9b, and correspond to the points* $\boldsymbol{\beta} = (1, 0)'$

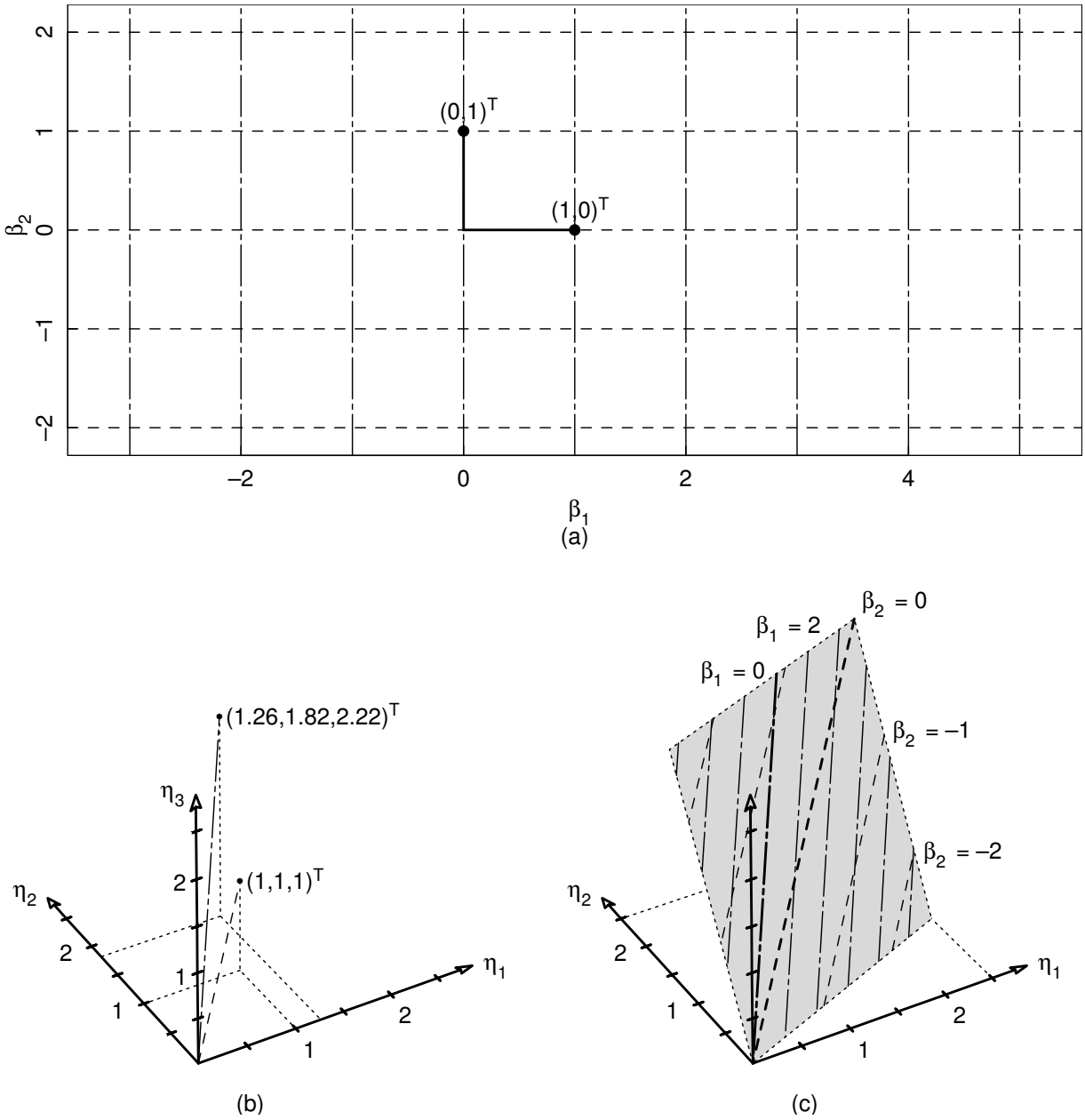

Figure 5.9: Expectation surface for the 3-case PCB example. Part a shows the parameter plane with β_1 parameter lines (dashed) and β_2 parameter lines (dot–dashed). Part b shows the vectors $\boldsymbol{x}_1$ (dashed line) and $\boldsymbol{x}_2$ (dot–dashed line) in the response space. The end points of the vectors correspond to $\boldsymbol{\beta} = (1, 0)'$ and $\boldsymbol{\beta} = (0, 1)'$ respectively. Part c shows a portion of the expectation plane (shaded) in the response space, with β_1 parameter lines (dashed) and β_2 parameter lines (dot–dashed).

and $\boldsymbol{\beta} = (0, 1)'$ in the parameter plane, shown in Figure 5.9a. The expectation function $\boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ defines a 2-dimensional expectation plane in the 3-dimensional response space. This is shown in Figure 5.9c, where the parameter lines corresponding to the lines $\beta_1 = -3, \ldots, 5$ and $\beta_2 = -2, \ldots, 2$, shown in Figure 5.9a, are given. A parameter line is associated with the parameter which is varying so the lines corresponding to $\beta_1 = -3, \ldots, 5$ (dotdashed lines) are called β_2 lines.

Note that the parameter lines in the parameter plane are straight, parallel, and equispaced, and that their images on the expectation plane are also straight, parallel, and equispaced. Because the vector $\boldsymbol{x}_1$ is shorter than $\boldsymbol{x}_2$ ($\|\boldsymbol{x}_1\| = \sqrt{3}$ while $\|\boldsymbol{x}_2\| = \sqrt{9.83}$), the spacing between the lines of constant β_1 on the expectation plane is less than that between the lines of constant β_2. Also, the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not orthogonal. The angle ω between them can be calculated from

$$\begin{aligned} \cos\omega &= \frac{\boldsymbol{x}_1'\boldsymbol{x}_2}{\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|} \\ &= \frac{5.30}{\sqrt{(3)(9.83)}} \\ &= 0.98 \end{aligned}$$

to be about 11°, so the parameter lines on the expectation plane are not at right angles as they are on the parameter plane.

As a consequence of the unequal length and nonorthogonality of the vectors, unit squares on the parameter plane map to parallelograms on the expectation plane. The area of the parallelogram is

$$\begin{aligned} \|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|\sin\omega &= \|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|\sqrt{1-\cos^2\omega} \\ &= \sqrt{(\boldsymbol{x}_1'\boldsymbol{x}_1)(\boldsymbol{x}_2'\boldsymbol{x}_2) - (\boldsymbol{x}_1'\boldsymbol{x}_2)^2} \\ &= \sqrt{|\boldsymbol{X}'\boldsymbol{X}|} \end{aligned} \tag{5.13}$$

That is, the Jacobian determinant *of the transformation from the parameter plane to the expectation plane is a constant equal to $|\boldsymbol{X}'\boldsymbol{X}|^{1/2}$. Conversely, the ratio of areas in the parameter plane to those on the expectation plane is $|\boldsymbol{X}'\boldsymbol{X}|^{-1/2}$.*

The simple linear mapping seen in the above example is true for all linear regression models. That is, for linear models, straight parallel equispaced lines in the parameter space map to straight parallel equispaced lines on the expectation plane in the response space. Consequently, rectangles in one plane map to parallelepipeds in the other plane, and circles or spheres in one plane map to ellipses or ellipsoids in the other plane. Furthermore, the Jacobian determinant, $|\boldsymbol{X}'\boldsymbol{X}|^{1/2}$, is a constant for linear models, and so regions of fixed size in one plane map to regions of fixed size in the other, no matter where they are on the plane. These properties, which make linear least squares especially simple, are discussed further in Section 1.2.3.

5.5.2 Determining the Least Squares Estimates

The geometric representation of linear least squares allows us to formulate a very simple scheme for determining the parameters estimates $\hat{\boldsymbol{\beta}}$. Since the expectation surface is linear, all we must

do to determine the point on the surface which is closest to the point $\boldsymbol{y}$, is to project $\boldsymbol{y}$ onto the expectation plane. This gives us $\hat{\boldsymbol{\eta}}$, and $\hat{\boldsymbol{\beta}}$ is then simply the value of $\boldsymbol{\beta}$ corresponding to $\hat{\boldsymbol{\eta}}$.

One approach to defining this projection is to observe that, after the projection, the residual vector $\boldsymbol{y} - \hat{\boldsymbol{\eta}}$ will be *orthogonal*, or *normal*, to the expectation plane. Equivalently, the residual vector must be orthogonal to all the columns of the $\boldsymbol{X}$ matrix, so

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$$

which is to say that the least squares estimate $\hat{\boldsymbol{\beta}}$ satisfies the *normal equations*

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \tag{5.14}$$

Because of (5.14) the least squares estimates are often written $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ as in (5.3). However, another way of expressing the estimate, and a more stable way of computing it, involves decomposing $\boldsymbol{X}$ into the product of an orthogonal matrix and an easily inverted matrix. Two such decompositions are the QR decomposition and the singular value decomposition (Dongarra, Bunch, Moler, and Stewart, 1979, Chapters 9 and 11). We use the QR decomposition, where

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$$

with the $N \times N$ matrix $\boldsymbol{Q}$ and the $N \times P$ matrix $\boldsymbol{R}$ constructed so that $\boldsymbol{Q}$ is orthogonal (that is, $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I}$) and $\boldsymbol{R}$ is zero below the main diagonal. Writing

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{Q} \end{bmatrix}$$

where $\boldsymbol{R}_1$ is $P \times P$ and upper triangular, and

$$\boldsymbol{Q} = [\boldsymbol{Q}_1 | \boldsymbol{Q}_2]$$

with $\boldsymbol{Q}_1$ the first P columns and $\boldsymbol{Q}_2$ the last $N - P$ columns of $\boldsymbol{Q}$, we have

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} = \boldsymbol{Q}_1\boldsymbol{R}_1 \tag{5.15}$$

Performing a QR decomposition is straightforward, as is shown in Appendix 2.

Geometrically, the columns of $\boldsymbol{Q}$ define an *orthonormal*, or *orthogonal*, basis for the response space with the property that the first P columns span the expectation plane. Projection onto the expectation plane is then very easy if we work in the coordinate system given by $\boldsymbol{Q}$. For example we transform the response vector to

$$\boldsymbol{w} = \boldsymbol{Q}'\boldsymbol{y} \tag{5.16}$$

with components

$$\boldsymbol{w}_1 = \boldsymbol{Q}_1'\boldsymbol{y} \tag{5.17}$$

and

$$\boldsymbol{w}_2 = \boldsymbol{Q}_2'\boldsymbol{y} \tag{5.18}$$

The projection of $\boldsymbol{w}$ onto the expectation plane is then simply

$$\begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{0} \end{bmatrix}$$

in the $\boldsymbol{Q}$ coordinates and

$$\hat{\boldsymbol{\eta}} = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_1 \boldsymbol{w}_1 \tag{5.19}$$

in the original coordinates.

Example 4. *As shown in Appendix 2, the* QR *decomposition (5.15) of the matrix*

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.26 \\ 1 & 1.82 \\ 1 & 2.22 \end{bmatrix}$$

for the 3-case PCB example is

$$\begin{bmatrix} 0.5774 & -0.7409 & 0.3432 \\ 0.5774 & 0.0732 & -0.8132 \\ 0.5774 & 0.6677 & 0.4700 \end{bmatrix} \begin{bmatrix} 1.7321 & 3.0600 \\ 0 & 0.6820 \\ 0 & 0 \end{bmatrix}$$

which gives [equation (5.16)]

$$\boldsymbol{w} = \begin{bmatrix} 3.23 \\ 1.16 \\ -0.24 \end{bmatrix}$$

In Figure 5.10a we show the expectation plane and observation vector in the original coordinate system. We also show the vectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3$, which are the columns of $\boldsymbol{Q}$. It can be seen that $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ lie in the expectation plane and $\boldsymbol{q}_3$ is orthogonal to it. In Figure 5.10b we show, in the transformed coordinates, the observation vector and the expectation plane, which is now horizontal. Note that projecting $\boldsymbol{w}$ onto the expectation plane is especially simple, since it merely requires replacing the last element in $\boldsymbol{w}$ by zero.

To determine the least squares estimate we must find the value $\hat{\boldsymbol{\beta}}$ corresponding to $\hat{\boldsymbol{\eta}}$. Since

$$\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

using (5.19) and (5.15)

$$\boldsymbol{R}_1 \hat{\boldsymbol{\beta}} = \boldsymbol{w}_1 \tag{5.20}$$

and we solve for $\hat{\boldsymbol{\beta}}$ by back-substitution Stewart (1973).

Example 5. *For the complete* ln(PCB), $\sqrt[3]{age}$ *data set,*

$$\boldsymbol{R}_1 = \begin{bmatrix} 5.29150 & 8.87105 \\ 0 & 2.16134 \end{bmatrix}$$

and $\boldsymbol{w}_1 = (7.7570, 4.9721)'$, *so* $\hat{\boldsymbol{\beta}} = (-2.391, 2.300)'$.

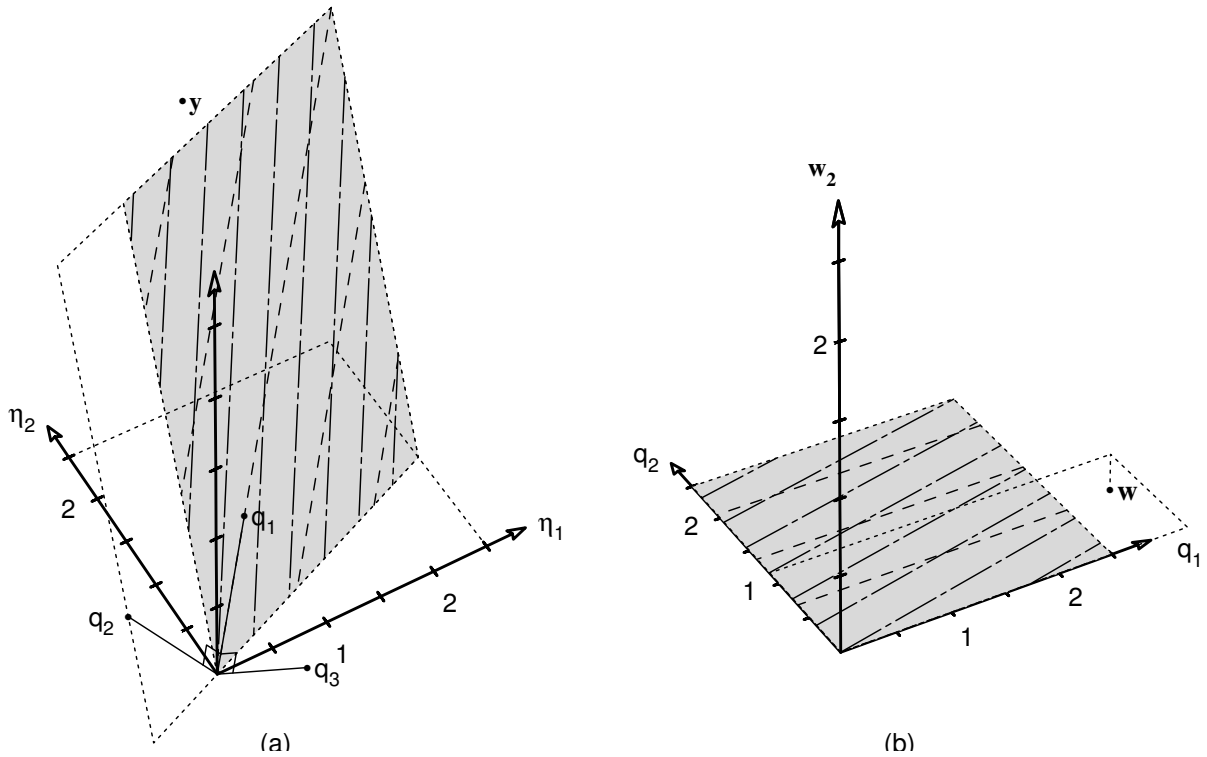

Figure 5.10: Expectation surface for the 3-case PCB example. Part a shows a portion of the expectation plane (shaded) in the response space with β_1 parameter lines (dashed) and β_2 parameter lines (dot–dashed) together with the response vector $\boldsymbol{y}$. Also shown are the orthogonal unit vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ in the expectation plane, and $\boldsymbol{q}_3$ orthogonal to the plane. Part b shows the response vector $\boldsymbol{w}$, and a portion of the expectation plane (shaded) in the rotated coordinates given by $\boldsymbol{Q}$.

5.5.3 Parameter Inference Regions

Just as the least squares estimates have informative geometric interpretations, so do the parameter inference regions (5.4), (5.5), (5.10) and those derived from (5.12). Such interpretations are helpful for understanding linear regression, and are essential for understanding nonlinear regression. (The geometric interpretation is less helpful in the Bayesian approach, so we discuss only the sampling theory and likelihood approaches.)

The main difference between the likelihood and sampling theory geometric interpretations is that the likelihood approach centers on the point $\boldsymbol{y}$ and the length of the residual vector at $\boldsymbol{\eta}(\boldsymbol{\beta})$ compared to the shortest residual vector, while the sampling theory approach focuses on possible values of $\boldsymbol{\eta}(\boldsymbol{\beta})$ and the angle that the resulting residual vectors could make with the expectation plane.

The Geometry of Sampling Theory Results

To develop the geometric basis of linear regression results from the sampling theory approach, we transform to the $\boldsymbol{Q}$ coordinate system. The model for the random variable $\mathcal{W} = \boldsymbol{Q}'\mathcal{Y}$ is

$$\mathcal{W} = \boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{Q}'\mathcal{Z}$$

or

$$\mathcal{U} = \mathcal{W} - \boldsymbol{R}\boldsymbol{\beta} \tag{5.21}$$

where $\mathcal{U} = \boldsymbol{Q}'\mathcal{Z}$.

The spherical normal distribution of $\mathcal{Z}$ is not affected by the orthogonal transformation, so $\mathcal{U}$ also has a spherical normal distribution. This can be established on the basis of the geometry, since the spherical probability contours will not be changed by a rigid rotation or reflection, which is what an orthogonal transformation must be. Alternatively, this can be established analytically because $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$, so the determinant of $\boldsymbol{Q}$ is ± 1 and $\|\boldsymbol{Q}\boldsymbol{x}\| = \|\boldsymbol{x}\|$ for any N-vector $\boldsymbol{x}$. Now the joint density for the random variables $\mathcal{Z} = (Z_1, \ldots, Z_n)'$ is

$$p_{\mathcal{Z}}(\boldsymbol{z}) = (2\pi\sigma^2)^{-N/2} \exp\left(\frac{-\boldsymbol{z}'\boldsymbol{z}}{2\sigma^2}\right)$$

and, after transformation, the joint density for $\mathcal{U} = \boldsymbol{Q}'\mathcal{Z}$ is

$$p_{\mathcal{U}}(\boldsymbol{u}) = (2\pi\sigma^2)^{-N/2}|\boldsymbol{Q}| \exp\left(\frac{-\boldsymbol{u}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{u}}{2\sigma^2}\right) = (2\pi\sigma^2)^{-N/2} \exp\left(\frac{-\boldsymbol{u}'\boldsymbol{u}}{2\sigma^2}\right)$$

From (5.21), the form of $\boldsymbol{R}$ leads us to partition $\mathcal{U}$ into two components:

$$\mathcal{U} = \begin{bmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \end{bmatrix}$$

where $\mathcal{U}_1$ consists of the first P elements of $\mathcal{U}$, and $\mathcal{U}_2$ the remaining $N - P$ elements. Each of these components has a spherical normal distribution of the appropriate dimension. Furthermore,

independence of elements in the original disturbance vector $\mathcal{Z}$ leads to independence of the elements of $\mathcal{U}$, so the components $\mathcal{U}_1$ and $\mathcal{U}_2$ are independent.

The dimensions ν_i of the components $\mathcal{U}_i$, called the *degrees of freedom*, are $\nu_1 = P$ and $\nu_2 = N - P$. The sum of squares of the coordinates of a ν-dimensional spherical normal vector has a $\sigma^2\chi^2$ distribution on ν degrees of freedom, so

$$\|\mathcal{U}_1\|^2 \sim \sigma^2\chi^2_P$$
$$\|\mathcal{U}_2\|^2 \sim \sigma^2\chi^2_{N-P}$$

where the symbol $\sim$ is read "is distributed as." Using the independence of $\mathcal{U}_1$ and $\mathcal{U}_2$, we have

$$\frac{\|\mathcal{U}_1\|^2/P}{\|\mathcal{U}_2\|^2/(N-P)} \sim F(P, N-P) \tag{5.22}$$

since the scaled ratio of two independent χ^2 random variables is distributed as Fisher's F distribution.

The distribution (5.22) gives a reference distribution for the ratio of the squared component lengths or, equivalently, for the angle that the disturbance vector makes with the horizontal plane. We may therefore use (5.21) and (5.22) to test the hypothesis that $\boldsymbol{\beta}$ equals some specific value, say $\boldsymbol{\beta}^0$, by calculating the residual vector $\boldsymbol{u}^0 = \boldsymbol{Q}'\boldsymbol{y} - \boldsymbol{R}\boldsymbol{\beta}^0$ and comparing the lengths of the components $\boldsymbol{u}^0_1$ and $\boldsymbol{u}^0_2$ as in (5.22). The reasoning here is that a large $\|\boldsymbol{u}^0_1\|$ compared to $\|\boldsymbol{u}^0_2\|$ suggests that the vector $\boldsymbol{y}$ is not very likely to have been generated by the model (**??**) with $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, since $\boldsymbol{u}^0$ has a suspiciously large component in the $\boldsymbol{Q}_1$ plane.

Note that

$$\frac{\|\boldsymbol{u}^0_2\|^2}{N-P} = \frac{S(\hat{\boldsymbol{\beta}})}{N-P} = s^2$$

and

$$\|\boldsymbol{u}^0_1\|^2 = \|\boldsymbol{R}_1\boldsymbol{\beta}^0 - \boldsymbol{w}_1\|^2 \tag{5.23}$$

and so the ratio (5.22) becomes

$$\frac{\|\boldsymbol{R}_1\boldsymbol{\beta}^0 - \boldsymbol{w}_1\|^2}{Ps^2} \tag{5.24}$$

Example 6. *We illustrate the decomposition of the residual* $\boldsymbol{u}$ *for testing the null hypothesis*

$$\mathrm{H}_0 : \boldsymbol{\beta} = (-2.0, 2.0)'$$

versus the alternative

$$\mathrm{H}_\mathrm{A} : \boldsymbol{\beta} \neq (-2.0, 2.0)'$$

for the full PCB data set in Figure 5.11. Even though the rotated data vector $\boldsymbol{w}$ *and the expectation surface for this example are in a 28-dimensional space, the relevant distances can be pictured in the 3-dimensional space spanned by the expectation surface (vectors* $\boldsymbol{q}_1$ *and* $\boldsymbol{q}_2$*) and the residual vector. The scaled lengths of the components* $\boldsymbol{u}_1$ *and* $\boldsymbol{u}_2$ *are compared to determine if the point* $\boldsymbol{\beta}^0 = (-2.0, 2.0)'$ *is reasonable.*

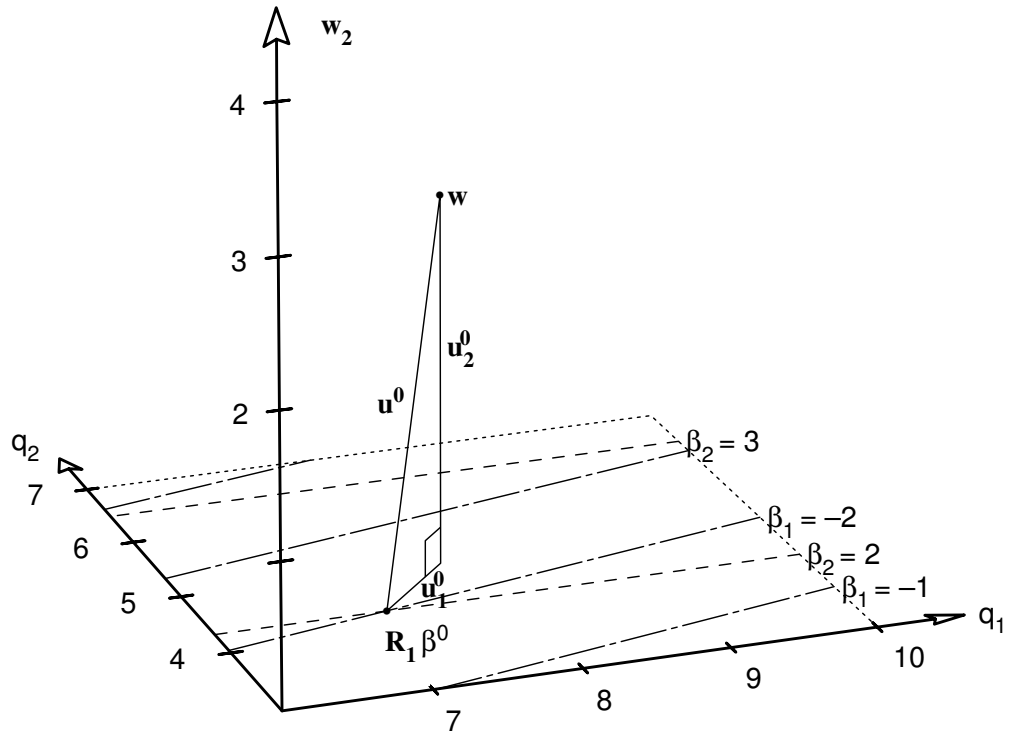

Figure 5.11: A geometric interpretation of the test $H_0 : \boldsymbol{\beta} = (-2.0, 2.0)'$ for the full PCB data set. We show the projections of the response vector $\boldsymbol{w}$ and a portion of the expectation plane projected into the 3-dimensional space given by the tangent vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, and the orthogonal component of the response vector, $\boldsymbol{w}_2$. For the test point $\boldsymbol{\beta}^0$, the residual vector $\boldsymbol{u}^0$ is decomposed into a tangential component $\boldsymbol{u}_1^0$ and an orthogonal component $\boldsymbol{u}_2^0$.

The numerator in (5.24) is

$$\left\| \begin{bmatrix} 5.29150 & 8.87105 \\ 0 & 2.16134 \end{bmatrix} \left[\begin{bmatrix} -2.0 \\ 2.0 \end{bmatrix} - \begin{bmatrix} 7.7570 \\ 4.9721 \end{bmatrix} \right] \right\|^2 = 0.882$$

The ratio is then $0.882/(2 \times 0.246) = 1.79$*, which corresponds to a tail probability (or p value) of 0.19 for an F distribution with 2 and 26 degrees of freedom. Since the probability of obtaining a ratio at least as large as 1.79 is 19%, we do not reject the null hypothesis.*

A $1-\alpha$ joint confidence region for the parameters $\boldsymbol{\beta}$ consists of all those values for which the above hypothesis test is not rejected at level α. Thus, a value $\boldsymbol{\beta}^0$ is within a $1-\alpha$ confidence region if

$$\frac{\|\boldsymbol{u}_1^0\|^2/P}{\|\boldsymbol{u}_2^0\|^2/(N-P)} \leq F_{P,N-P;\alpha}$$

Since s^2 does not depend on $\boldsymbol{\beta}^0$, the points inside the confidence region form a disk on the expectation plane defined by

$$\|\boldsymbol{u}_1\|^2 \leq P s^2 F_{P,N-P;\alpha}$$

Furthermore, from (5.20) and (5.23) we have

$$\|\boldsymbol{u}_1\|^2 = \|\boldsymbol{R}_1(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|^2$$

so a point on the boundary of the confidence region in the parameter space satisfies

$$\boldsymbol{R}_1(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \sqrt{P s^2 F_{P,N-P;\alpha}}\, \boldsymbol{d}$$

where $\|\boldsymbol{d}\| = 1$. That is, the confidence region is given by

$$\left\{ \boldsymbol{\beta} = \hat{\boldsymbol{\beta}} + \sqrt{P s^2 F_{P,N-P;\alpha}} \boldsymbol{R}_1^{-1} \boldsymbol{d} \mid \|\boldsymbol{d}\| = 1 \right\} \tag{5.25}$$

Thus the region of "reasonable" parameter values is a disk centered at $\boldsymbol{R}_1\hat{\boldsymbol{\beta}}$ on the expectation plane and is an ellipse centered at $\hat{\boldsymbol{\beta}}$ in the parameter space.

Example 7. *For the* ln(PCB) *versus* $\sqrt[3]{age}$ *data,* $\hat{\boldsymbol{\beta}} = (-2.391, 2.300)'$ *and* $s^2 = 0.246$ *based on* 26 *degrees of freedom, so the 95% confidence disk on the transformed expectation surface is*

$$\boldsymbol{R}_1\boldsymbol{\beta} = \begin{bmatrix} 7.7570 \\ 4.9721 \end{bmatrix} + 1.288 \begin{bmatrix} \cos\omega \\ \sin\omega \end{bmatrix}$$

where $0 \leq \omega \leq 2\pi$*. The disk is shown in the expectation plane in Figure 5.12a, and the corresponding ellipse*

$$\boldsymbol{\beta} = \begin{bmatrix} -2.391 \\ 2.300 \end{bmatrix} + 1.288 \begin{bmatrix} 0.18898 & -0.77566 \\ 0 & 0.46268 \end{bmatrix} \begin{bmatrix} \cos\omega \\ \sin\omega \end{bmatrix}$$

is shown in the parameter plane in Figure 5.12b.

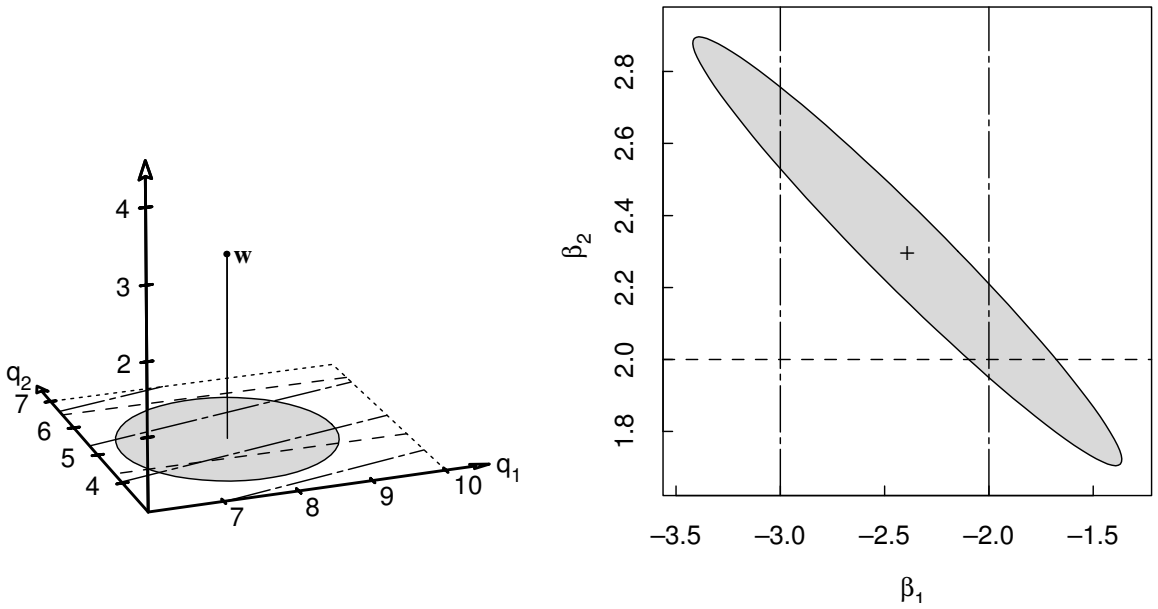

Figure 5.12: The 95% confidence disk and parameter confidence region for the PCB data. Part *a* shows the response vector $\boldsymbol{w}$ and a portion of the expectation plane projected into the 3-dimensional space given by the tangent vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, and the orthogonal component of the response vector, $\boldsymbol{w}_2$. The 95% confidence disk (shaded) in the expectation plane (part *a*) maps to the elliptical confidence region (shaded) in the parameter plane (part *b*).

5.5.4 Marginal Confidence Intervals

We can create a marginal confidence interval for a single parameter, say β_1, by "inverting" a hypothesis test of the form

$$\mathrm{H}_0 : \beta_1 = \beta_1^0$$

versus

$$\mathrm{H}_\mathrm{A} : \beta_1 \neq \beta_1^0$$

Any β_1^0 for which H_0 is not rejected at level α is included in the $1-\alpha$ confidence interval. To perform the hypothesis test, we choose any parameter vector with $\beta_1 = \beta_1^0$, say $\left(\beta_1^0, \mathbf{0}'\right)'$, calculate the transformed residual vector $\boldsymbol{u}^0$, and divide it into three components: the first component $\boldsymbol{u}_1^0$ of dimension $P-1$ and parallel to the hyperplane defined by $\beta_1 = \beta_1^0$; the second component u_2^0 of dimension 1 and in the expectation plane but orthogonal to the β_1^0 hyperplane; and the third component $\boldsymbol{u}_3^0$ of length $(N-P)s^2$ and orthogonal to the expectation plane. The component u_2^0 is the same for any parameter $\boldsymbol{\beta}$ with $\beta_1 = \beta_1^0$, and, assuming that the true β_1 is β_1^0, the scaled ratio of the corresponding random variables U_2 and $\boldsymbol{U}_3$ has the distribution

$$\frac{\mathcal{U}_2^2/1}{\|\mathcal{U}_3\|^2/(N-P)} \sim F(1, N-P)$$

Thus we reject H_0 at level α if

$$\left(u_2^0\right)^2 > s^2 F(1, N-P; \alpha)$$

Example 8. *To test the null hypothesis*

$$\mathrm{H}_0 : \beta_1 = -2.0$$

versus the alternative

$$\mathrm{H}_\mathrm{A} : \beta_1 \neq -2.0$$

for the complete PCB data set, we decompose the transformed residual vector at $\boldsymbol{\beta}^0 = (-2.0, 2.2)'$ *into three components as shown in Figure 5.13 and calculate the ratio*

$$\frac{\left(u_2^0\right)^2}{s^2} = \frac{0.240}{0.246} = 0.97$$

This corresponds to a p value of 0.33, and so we do not reject the null hypothesis.

We can create a $1-\alpha$ marginal confidence interval for β_1 as all values for which

$$\left(u_2^0\right)^2 \leq s^2 F(1, N-P; \alpha)$$

or, equivalently,

$$|u_2^0| \leq s \cdot t_{N-P;\alpha} \tag{5.26}$$

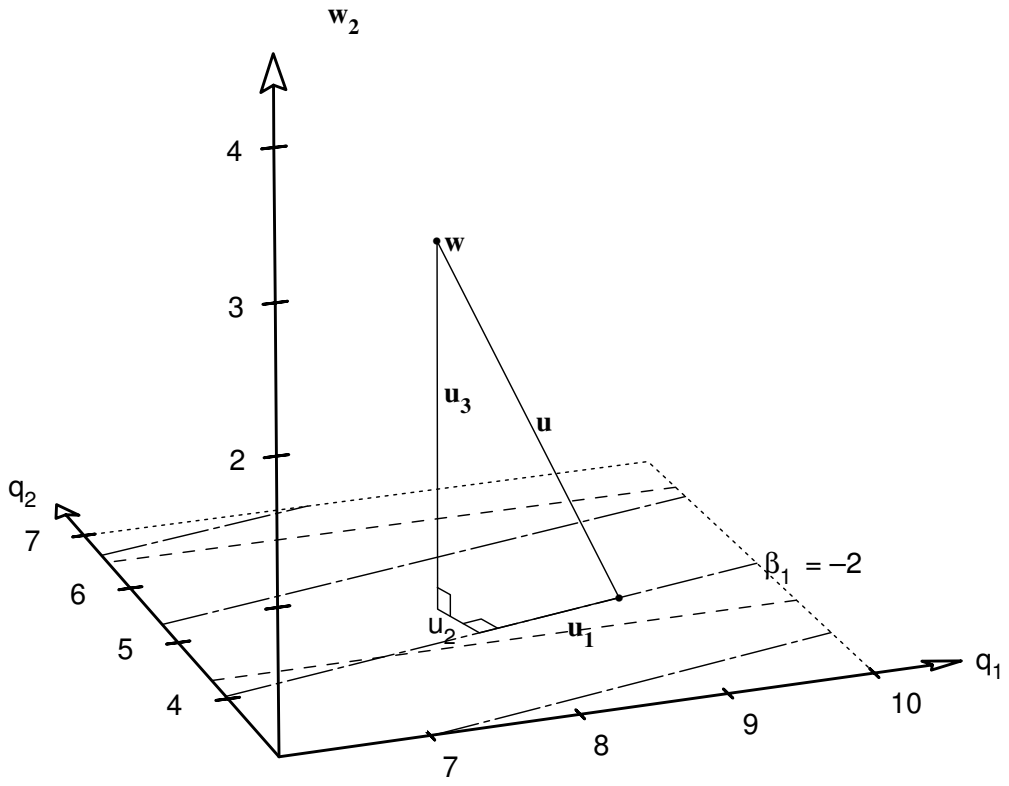

Figure 5.13: A geometric interpretation of the test H_0:$\beta_1 = -2.0$ for the full PCB data set. We show the response vector $\boldsymbol{w}$, and a portion of the expectation plane projected into the 3-dimensional space given by the tangent vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, and the orthogonal component of the response vector, $\boldsymbol{w}_2$. For a representative point on the line $\beta_1 = -2$ the residual vector $\boldsymbol{u}$ is decomposed into a tangential component $\boldsymbol{u}_1^0$ along the line, a tangential component u_2^0 perpendicular to the line, and an orthogonal component $\boldsymbol{u}_3^0$.

Since $|u_2^0|$ is the distance from the point $\boldsymbol{R}_1\hat{\boldsymbol{\beta}}$ to the line corresponding to $\beta_1 = \beta_1^0$ on the transformed parameter plane, the confidence interval will include all values β_1^0 for which the corresponding parameter line intersects the disk

$$\left\{\boldsymbol{R}_1\hat{\boldsymbol{\beta}} + st_{N-P;\alpha}\boldsymbol{d} \mid \|\boldsymbol{d}\| = 1\right\} \tag{5.27}$$

Instead of determining the value of $|u_2^0|$ for each β_1^0, we take the disk (5.27) and determine the minimum and maximum values of β_1 for points on the disk. Writing $\boldsymbol{r}^1$ for the first row of $\boldsymbol{R}_1^{-1}$, the values of β_1 corresponding to points on the expectation plane disk are

$$\boldsymbol{r}^1(\boldsymbol{R}_1\hat{\boldsymbol{\beta}} + s \cdot t_{N-P;\alpha}\boldsymbol{d}) = \hat{\beta}_1 + s \cdot t_{N-P;\alpha}\boldsymbol{r}^1\boldsymbol{d}$$

and the minimum and maximum occur for the unit vectors in the direction of $\boldsymbol{r}^1$; that is, $\boldsymbol{d} = \pm\left(\boldsymbol{r}^1\right)'/\|\boldsymbol{r}^1\|$. This gives the confidence interval

$$\hat{\beta}_1 \pm s\|\boldsymbol{r}^1\|t_{N-P;\alpha}$$

In general, a marginal confidence interval for parameter β_p is

$$\hat{\beta}_p \pm s\|\boldsymbol{r}^p\|t_{N-P;\alpha} \tag{5.28}$$

where $\boldsymbol{r}^p$ is the pth row of $\boldsymbol{R}_1^{-1}$. The quantity

$$\text{se}(\hat{\beta}_p) = s\|\boldsymbol{r}^p\| \tag{5.29}$$

is called the *standard error* for the pth parameter. Since

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{R}_1'\boldsymbol{R}_1)^{-1} = \boldsymbol{R}_1^{-1}\boldsymbol{R}_1^{-\mathrm{T}}$$

$\|\boldsymbol{r}^p\|^2 = \left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}_{pp}$, so the standard error can be written as in equation (5.6).

A convenient summary of the variability of the parameter estimates can be obtained by factoring $\boldsymbol{R}_1^{-1}$ as

$$\boldsymbol{R}_1^{-1} = \text{diag}(\|\boldsymbol{r}^1\|, \|\boldsymbol{r}^2\|, \ldots, \|\boldsymbol{r}^P\|)\boldsymbol{L} \tag{5.30}$$

where $\boldsymbol{L}$ has unit length rows. The diagonal matrix provides the parameter standard errors, while the *correlation matrix*

$$\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}' \tag{5.31}$$

gives the correlations between the parameter estimates.

Example 9. *For the* ln(PCB) *data,* $\hat{\boldsymbol{\beta}} = (-2.391, 2.300)'$, $s^2 = 0.246$ *with 26 degrees of freedom, and*

$$\begin{aligned}
\boldsymbol{R}_1^{-1} &= \begin{bmatrix} 5.29150 & 8.87105 \\ 0 & 2.16134 \end{bmatrix}^{-1} \\
&= \begin{bmatrix} 0.18898 & -0.77566 \\ 0 & 0.46268 \end{bmatrix} \\
&= \begin{bmatrix} 0.798 & 0 \\ 0 & 0.463 \end{bmatrix} \begin{bmatrix} 0.237 & -0.972 \\ 0 & 1 \end{bmatrix}
\end{aligned}$$

which gives standard errors of $0.798\sqrt{0.246} = 0.396$ *for* β_1 *and* $0.463\sqrt{0.246} = 0.230$ *for* β_2*. Also*

$$\boldsymbol{C} = \begin{bmatrix} 1 & -0.972 \\ -0.972 & 1 \end{bmatrix}$$

so the correlation between β_1 *and* β_2 *is* -0.97*. The 95% confidence intervals for the parameters are given by* $-2.391 \pm 2.056(0.396)$ *and* $2.300 \pm 2.056(0.230)$*, which are plotted in Figure 5.8a.*

Marginal confidence intervals for the expected response at a design point $\boldsymbol{x}_0$ can be created by determining which hyperplanes formed by constant $\boldsymbol{x}_0'\boldsymbol{\beta}$ intersect the disk (5.27). Using the same argument as was used to derive (5.28), we obtain a standard error for the expected response at $\boldsymbol{x}_0$ as $s\|\boldsymbol{x}_0'\boldsymbol{R}_1^{-1}\|$, so the confidence interval is

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm s\|\boldsymbol{x}_0'\boldsymbol{R}_1^{-1}\|t_{N-P;\alpha} \tag{5.32}$$

Similarly, a confidence band for the response function is

$$\boldsymbol{x}'\hat{\boldsymbol{\beta}} \pm s\|\boldsymbol{x}'\boldsymbol{R}_1^{-1}\|\sqrt{PF_{P,N-P;\alpha}} \tag{5.33}$$

Example 10. *A plot of the fitted expectation function and the 95% confidence bands for the PCB example was given in Figure 5.8b.*

Bibliography

C. A. Bache, J. W. Serum, W. D. Youngs, and D. J. Lisk. Polychlorinated Biphenyl residues: Accumulation in Cayuga Lake trout with age. *Science*, 117:1192–1193, 1972.

George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA, 1973.

George E. P. Box and Paul W. Tidwell. Transformations of the independent variables. *Technometrics*, 4:531–550, 1962.

J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart. *Linpack Users' Guide*. SIAM, Philadelphia, 1979.

G. A. F. Seber. *Linear Regression Analysis*. Wiley, New York, 1977.

G. W. Stewart. *Introduction to Matrix Computations*. Academic Press, New York, 1973.