Chapter 3

Inference

Up till now, we haven't found it necessary to assume any distributional form for the errors ε. However, if we want to make any confidence intervals or perform any hypothesis tests, we will need to do this. The usual assumption is that the errors are normally distributed and in practice this is often, although not always, a reasonable assumption. We'll assume that the errors are independent and identically normally distributed with mean 0 and variance σ^2, i.e.

$$\varepsilon \sim N(0, \sigma^2 I)$$

We can handle non-identity variance matrices provided we know the form — see the section on generalized least squares later. Now since $y = X\beta + \varepsilon$,

$$y \sim N(X\beta, \sigma^2 I)$$

is a compact description of the regression model and from this we find that (using the fact that linear combinations of normally distributed values are also normal)

$$\hat{\beta} = (X^T X)^{-1} X^T y \sim N(\beta, (X^T X)^{-1} \sigma^2)$$

3.1 Hypothesis tests to compare models

Given several predictors for a response, we might wonder whether all are needed. Consider a large model, Ω, and a smaller model, ω, which consists of a subset of the predictors that are in Ω. By the principle of Occam's Razor (also known as the law of parsimony), we'd prefer to use ω if the data will support it. So we'll take ω to represent the null hypothesis and Ω to represent the alternative. A geometric view of the problem may be seen in Figure 3.1.

If $\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega$ is small, then ω is an adequate model relative to Ω. This suggests that something like

$$\frac{\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega}{\mathrm{RSS}_\Omega}$$

would be a potentially good test statistic where the denominator is used for scaling purposes.

As it happens the same test statistic arises from the likelihood-ratio testing approach. We give an outline of the development: If $L(\beta, \sigma|y)$ is likelihood function, then the likelihood ratio statistic is

$$\frac{\max_{\beta,\sigma\in\Omega} L(\beta, \sigma|y)}{\max_{\beta,\sigma\in\omega} L(\beta, \sigma|y)}$$

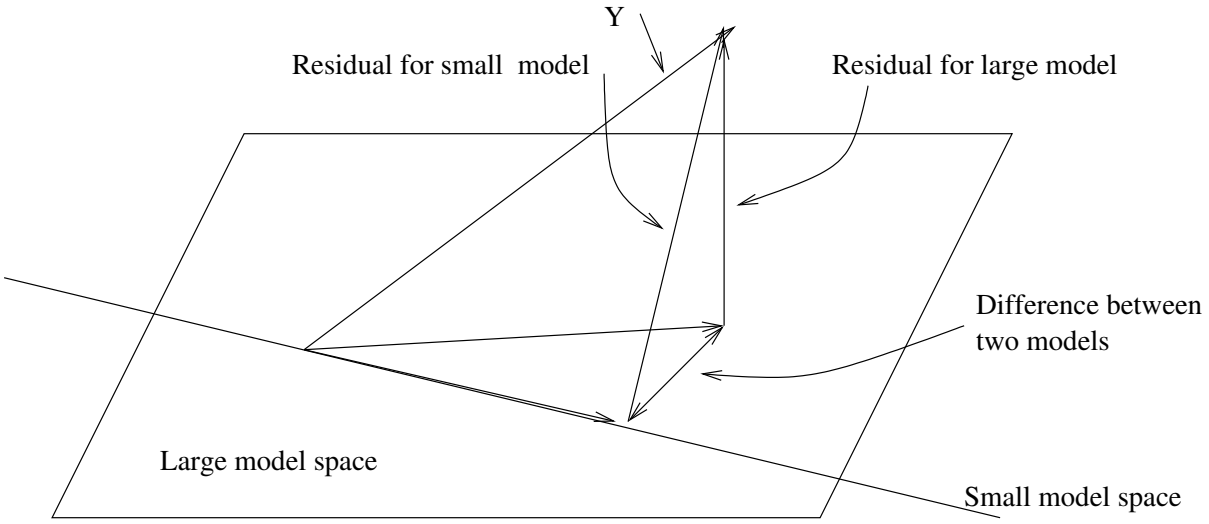

Figure 3.1: Geometric view of the comparison between big model, Ω, and small model, ω. The squared length of the residual vector for the big model is RSS_Ω while that for the small model is RSS_ω. By Pythagoras' theorem, the squared length of the vector connecting the two fits is $RSS_\omega - RSS_\Omega$. A small value for this indicates that the small model fits almost as well as the large model and thus might be preferred due to its simplicity.

The test should reject if this ratio is too large. Working through the details, we find that

$$L(\hat{\beta}, \hat{\sigma}|y) \propto \hat{\sigma}^{-n}$$

which gives us a test that rejects if

$$\frac{\hat{\sigma}^2_\omega}{\hat{\sigma}^2_\Omega} > \text{a constant}$$

which is equivalent to

$$\frac{\text{RSS}_\omega}{\text{RSS}_\Omega} > \text{a constant}$$

(constants are not the same) or

$$\frac{\text{RSS}_\omega}{\text{RSS}_\Omega} - 1 > \text{a constant} - 1$$

which is

$$\frac{\text{RSS}_\omega - \text{RSS}_\Omega}{\text{RSS}_\Omega} > \text{a constant}$$

which is the same statistics suggested by the geometric view. It remains for us to discover the null distribution of this statistic.

Now suppose that the dimension (no. of parameters) of Ω is q and dimension of ω is p. Now by Cochran's theorem, if the null (ω) is true then

$$\frac{\text{RSS}_\omega - \text{RSS}_\Omega}{q-p} \sim \sigma^2 \chi^2_{q-p} \qquad \frac{\text{RSS}_\Omega}{n-q} \sim \sigma^2 \chi^2_{n-q}$$

and these two quantities are independent. So we find that

$$F = \frac{(\text{RSS}_\omega - \text{RSS}_\Omega)/(q-p)}{\text{RSS}_\Omega/(n-q)} \sim F_{q-p,n-q}$$

Thus we would reject the null hypothesis if $F > F^{(\alpha)}_{q-p,n-q}$ The degrees of freedom of a model is (usually) the number of observations minus the number of parameters so this test statistic can be written

$$F = \frac{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/(df_\omega - df_\Omega)}{\mathrm{RSS}_\Omega/df_\Omega}$$

where $df_\Omega = n - q$ and $df_\omega = n - p$. The same test statistic applies not just when ω is a subset of Ω but also to a subspace. This test is very widely used in regression and analysis of variance. When it is applied in different situations, the form of test statistic may be re-expressed in various different ways. The beauty of this approach is you only need to know the general form. In any particular case, you just need to figure out which models represents the null and alternative hypotheses, fit them and compute the test statistic. It is very versatile.

3.2 Some Examples

3.2.1 Test of all predictors

Are any of the predictors useful in predicting the response?

- Full model (Ω) : $y = X\beta + \varepsilon$ where X is a full-rank $n \times p$ matrix.
- Reduced model (ω) : $y = \mu + \varepsilon$ — predict y by the mean.

We could write the null hypothesis in this case as

$$H_0 : \beta_1 = \ldots \beta_{p-1} = 0$$

Now

- $\mathrm{RSS}_\Omega = (y - X\hat{\beta})^T(y - X\hat{\beta}) = \hat{\varepsilon}^T\hat{\varepsilon} = \mathrm{RSS}$
- $\mathrm{RSS}_\omega = (y - \bar{y})^T(y - \bar{y})$ =SYY, which is sometimes known as the sum of squares corrected for the mean.

So in this case

$$F = \frac{(\mathrm{SYY} - \mathrm{RSS})/(p-1)}{\mathrm{RSS}/(n-p)}$$

We'd now refer to $F_{p-1,n-p}$ for a critical value or a p-value. Large values of F would indicate rejection of the null. Traditionally, the information in the above test is presented in an *analysis of variance table*. Most computer packages produce a variant on this. See Table 3.1. It is not really necessary to specifically compute all the elements of the table. As the originator of the table, Fisher said in 1931, it is "nothing but a convenient way of arranging the arithmetic". Since he had to do his calculations by hand, the table served some purpose but it is less useful now.

A failure to reject the null hypothesis is not the end of the game — you must still investigate the possibility of non-linear transformations of the variables and of outliers which may obscure the relationship. Even then, you may just have insufficient data to demonstrate a real effect which is why we must be careful to say "fail to reject" the null rather than "accept" the null. It would be a mistake to conclude that no real relationship exists. This issue arises when a pharmaceutical company wishes to show that a proposed generic replacement for a brand-named drug is equivalent. It would not be enough in this instance just to fail to reject the null. A higher standard would be required.

Source	Deg. of Freedom	Sum of Squares	Mean Square	F
Regression	$p-1$	SS_{reg}	$SS_{reg}/(p-1)$	F
Residual	n-p	RSS	$RSS/(n-p)$	
Total	n-1	SYY		

Table 3.1: Analysis of Variance table

When the null is rejected, this does not imply that the alternative model is the best model. We don't know whether all the predictors are required to predict the response or just some of them. Other predictors might also be added — for example quadratic terms in the existing predictors. Either way, the overall F-test is just the beginning of an analysis and not the end.

Let's illustrate this test and others using an old economic dataset on 50 different countries. These data are averages over 1960-1970 (to remove business cycle or other short-term fluctuations). `dpi` is per-capita disposable income in U.S. dollars; `ddpi` is the percent rate of change in per capita disposable income; `sr` is aggregate personal saving divided by disposable income. The percentage population under 15 (`pop15`) and over 75 (`pop75`) are also recorded. The data come from Belsley, Kuh, and Welsch (1980). Take a look at the data:

```
> data(savings)
> savings
                sr pop15 pop75     dpi  ddpi
Australia      11.43 29.35  2.87 2329.68  2.87
Austria        12.07 23.32  4.41 1507.99  3.93
--- cases deleted ---
Malaysia        4.71 47.20  0.66  242.69  5.08
```

First consider a model with all the predictors:

```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
> summary(g)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.566087   7.354516    3.88  0.00033
pop15       -0.461193   0.144642   -3.19  0.00260
pop75       -1.691498   1.083599   -1.56  0.12553
dpi         -0.000337   0.000931   -0.36  0.71917
ddpi         0.409695   0.196197    2.09  0.04247

Residual standard error: 3.8 on 45 degrees of freedom
Multiple R-Squared: 0.338,      Adjusted R-squared: 0.28
F-statistic: 5.76 on 4 and 45 degrees of freedom,       p-value: 0.00079
```

We can see directly the result of the test of whether any of the predictors have significance in the model. In other words, whether $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. Since the p-value is so small, this null hypothesis is rejected.

We can also do it directly using the F-testing formula:

```
> sum((savings$sr-mean(savings$sr))^2)
[1] 983.63
> sum(g$res^2)
[1] 650.71
> ((983.63-650.71)/4)/(650.706/45)
[1] 5.7558
> 1-pf(5.7558,4,45)
[1] 0.00079026
```

Do you know where all the numbers come from? Check that they match the regression summary above.

3.2.2 Testing just one predictor

Can one particular predictor be dropped from the model? The null hypothesis would be $H_0 : \beta_i = 0$. Set it up like this

- RSS_Ω is the RSS for the model with all the predictors of interest (p parameters).
- RSS_ω is the RSS for the model with all the above predictors except predictor i.

The F-statistic may be computed using the formula from above. An alternative approach is to use a t-statistic for testing the hypothesis:

$$t_i = \hat{\beta}_i / se(\hat{\beta}_i)$$

and check for significance using a t distribution with $n - p$ degrees of freedom.

However, squaring the t-statistic here, i.e. t_i^2 gives you the F-statistic, so the two approaches are identical.

For example, to test the null hypothesis that $\beta_1 = 0$ i.e. that `p15` is not significant in the full model, we can simply observe that the p-value is 0.0026 from the table and conclude that the null should be rejected.

Let's do the same test using the general F-testing approach: We'll need the RSS and df for the full model — these are 650.71 and 45 respectively.

and then fit the model that represents the null:

```
> g2 <- lm(sr ~ pop75 + dpi + ddpi, data=savings)
```

and compute the RSS and the F-statistic:

```
> sum(g2$res^2)
[1] 797.72
> (797.72-650.71)/(650.71/45)
[1] 10.167
```

The p-value is then

```
> 1-pf(10.167,1,45)
[1] 0.0026026
```

We can relate this to the t-based test and p-value by

```
> sqrt(10.167)
[1] 3.1886
> 2*(1-pt(3.1886,45))
[1] 0.0026024
```

A somewhat more convenient way to compare two nested models is

```
> anova(g2,g)
Analysis of Variance Table

Model 1: sr ~ pop75 + dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     46        798
2     45        651  1    147    10.2 0.0026
```

Understand that this test of `pop15` is relative to the other predictors in the model, namely `pop75, dpi` and `ddpi`. If these other predictors were changed, the result of the test may be different. This means that it is not possible to look at the effect of `pop15` in isolation. Simply stating the null hypothesis as $H_0 : \beta_{pop15} = 0$ is insufficient — information about what other predictors are included in the null is necessary. The result of the test may be different if the predictors change.

3.2.3 Testing a pair of predictors

Suppose we wish to test the significance of variables X_j and X_k. We might construct a table as shown just above and find that both variables have p-values greater than 0.05 thus indicating that individually neither is significant. Does this mean that both X_j and X_k can be eliminated from the model? *Not necessarily*

Except in special circumstances, dropping one variable from a regression model causes the estimates of the other parameters to change so that we might find that after dropping X_j, that a test of the significance of X_k shows that it should now be included in the model.

If you really want to check the joint significance of X_j and X_k, you should fit a model with and then without them and use the general F-test discussed above. Remember that even the result of this test may depend on what other predictors are in the model.

Can you see how to test the hypothesis that both `pop75` and `ddpi` may be excluded from the model?

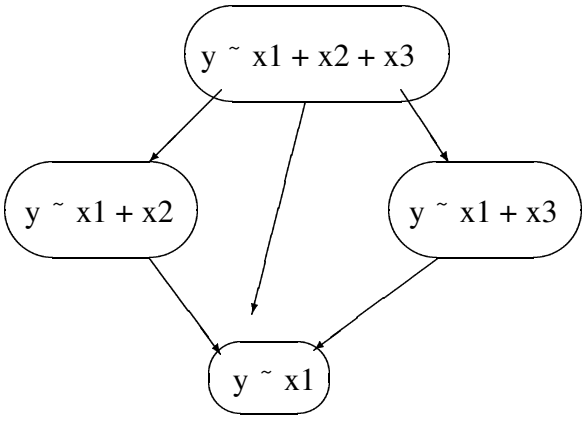

Figure 3.2: Testing two predictors

The testing choices are depicted in Figure 3.2. Here we are considering two predictors, `x2` and `x3` in the presence of `x1`. Five possible tests may be considered here and the results may not always be apparently consistent. The results of each test need to be considered individually in the context of the particular example.

3.2.4 Testing a subspace

Consider this example. Suppose that y is the miles-per-gallon for a make of car and X_j is the weight of the engine and X_k is the weight of the rest of the car. There would also be some other predictors. We might wonder whether we need two weight variables — perhaps they can be replaced by the total weight, $X_j + X_k$. So if the original model was

$$y = \beta_0 + \ldots + \beta_j X_j + \beta_k X_k + \ldots + \varepsilon$$

then the reduced model is

$$y = \beta_0 + \ldots + \beta_l (X_j + X_k) + \ldots + \varepsilon$$

which requires that $\beta_j = \beta_k$ for this reduction to be possible. So the null hypothesis is

$$H_0 : \beta_j = \beta_k$$

This defines a linear subspace to which the general F-testing procedure applies. In our example, we might hypothesize that the effect of young and old people on the savings rate was the same or in other words that

$$H_0 : \beta_{pop15} = \beta_{pop75}$$

In this case the null model would take the form

$$y = \beta_0 + \beta_{pop15}(pop15 + pop75) + \beta_{dpi} dpi + \beta_{ddpi} ddpi + \varepsilon$$

We can then compare this to the full model as follows:

```
> g <- lm(sr ~ .,savings)
> gr <- lm(sr ~ I(pop15+pop75)+dpi+ddpi,savings)
> anova(gr,g)
Analysis of Variance Table

Model 1: sr ~ I(pop15 + pop75) + dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     46        674
2     45        651  1     23    1.58   0.21
```

The period in the first model formula is short hand for all the other variables in the data frame. The function `I()` ensures that the argument is evaluated rather than interpreted as part of the model formula. The p-value of 0.21 indicates that the null cannot be rejected here meaning that there is not evidence here that young and old people need to be treated separately in the context of this particular model.

Suppose we want to test whether one of the coefficients can be set to a particular value. For example,

$$H_0 : \beta_{ddpi} = 1$$

Here the null model would take the form:

$$y = \beta_0 + \beta_{pop15} pop15 + \beta_{pop75} pop75 + \beta_{dpi} dpi + ddpi + \varepsilon$$

Notice that there is now no coefficient on the `ddpi` term. Such a fixed term in the regression equation is called an *offset*. We fit this model and compare it to the full:

```
> gr <- lm(sr ~ pop15+pop75+dpi+offset(ddpi),savings)
> anova(gr,g)
Analysis of Variance Table

Model 1: sr ~ pop15 + pop75 + dpi + offset(ddpi)
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     46        782
2     45        651  1    131    9.05 0.0043
```

We see that the p-value is small and the null hypothesis here is soundly rejected. A simpler way to test such point hypotheses is to use a t-statistic:

$$t = (\hat{\beta} - c)/\text{se}(\hat{\beta})$$

where c is the point hypothesis. So in our example the statistic and corresponding p-value is

```
> tstat <- (0.409695-1)/0.196197
> tstat
[1] -3.0087
> 2*pt(tstat,45)
[1] 0.0042861
```

We can see the p-value is the same as before and if we square the t-statistic

```
> tstat^2
[1] 9.0525
```

we find we get the F-value. This latter approach is preferred in practice since we don't need to fit two models but it is important to understand that it is equivalent to the result obtained using the general F-testing approach.

Can we test a hypothesis such as

$$H_0 : \beta_j \beta_k = 1$$

using our general theory?

No. This hypothesis is not linear in the parameters so we can't use our general method. We'd need to fit a non-linear model and that lies beyond the scope of this book.

3.3 Concerns about Hypothesis Testing

1. The general theory of hypothesis testing posits a *population* from which a *sample* is drawn — this is our data. We want to say something about the unknown *population* values β using estimated values $\hat{\beta}$ that are obtained from the *sample* data. Furthermore, we require that the data be generated using a *simple random sample* of the population. This sample is finite in size, while the population is infinite in size or at least so large that the sample size is a negligible proportion of the whole. For more complex sampling designs, other procedures should be applied, but of greater concern is the case when the data is not a random sample at all. There are two cases:

   (a) A sample of convenience is where the data is not collected according to a sampling design. In some cases, it may be reasonable to proceed as if the data were collected using a random mechanism. For example, suppose we take the first 400 people from the phonebook whose

names begin with the letter P. Provided there is no ethnic effect, it may be reasonable to consider this a random sample from the population defined by the entries in the phonebook. Here we are assuming the selection mechanism is effectively random with respect to the objectives of the study. An assessment of *exchangeability* is required - are the data as good as random? Other situations are less clear cut and judgment will be required. Such judgments are easy targets for criticism. Suppose you are studying the behavior of alcoholics and advertise in the media for study subjects. It seems very likely that such a sample will be biased perhaps in unpredictable ways. In cases such as this, a sample of convenience is clearly biased in which case conclusions must be limited to the sample itself. This situation reduces to the next case, where the sample is the population.

Sometimes, researchers may try to select a "representative" sample by hand. Quite apart from the obvious difficulties in doing this, the logic behind the statistical inference depends on the sample being random. This is not to say that such studies are worthless but that it would be unreasonable to apply anything more than descriptive statistical techniques. Confidence in the of conclusions from such data is necessarily suspect.

(b) The sample is the complete population in which case one might argue that inference is not required since the population and sample values are one and the same. For both regression datasets we have considered so far, the sample is effectively the population or a large and biased proportion thereof.

In these situations, we can put a different meaning to the hypothesis tests we are making. For the Galapagos dataset, we might suppose that if the number of species had no relation to the five geographic variables, then the observed response values would be randomly distributed between the islands without relation to the predictors. We might then ask what the chance would be under this assumption that an F-statistic would be observed as large or larger than one we actually observed. We could compute this exactly by computing the F-statistic for all possible (30!) permutations of the response variable and see what proportion exceed the observed F-statistic. This is a permutation test. If the observed proportion is small, then we must reject the contention that the response is unrelated to the predictors. Curiously, this proportion is estimated by the p-value calculated in the usual way based on the assumption of normal errors thus saving us from the massive task of actually computing the regression on all those computations.

Let see how we can apply the permutation test to the savings data. I chose a model with just `pop75` and `dpi` so as to get a p-value for the F-statistic that is not too small.

```
> g <- lm(sr ~ pop75+dpi,data=savings)
> summary(g)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.056619   1.290435    5.47  1.7e-06
pop75        1.304965   0.777533    1.68     0.10
dpi         -0.000341   0.001013   -0.34     0.74

Residual standard error: 4.33 on 47 degrees of freedom
Multiple R-Squared: 0.102,      Adjusted R-squared: 0.0642
F-statistic: 2.68 on 2 and 47 degrees of freedom, p-value: 0.0791
```

We can extract the F-statistic as

```
> gs <- summary(g)
```

```
> gs$fstat
  value   numdf   dendf
 2.6796  2.0000 47.0000
```

The function `sample()` generates random permutations. We compute the F-statistic for 1000 randomly selected permutations and see what proportion exceed the the F-statistic for the original data:

```
> fstats <- numeric(1000)
> for(i in 1:1000){
+ ge <- lm(sample(sr) ~ pop75+dpi,data=savings)
+ fstats[i] <- summary(ge)$fstat[1]
+ }
> length(fstats[fstats > 2.6796])/1000
[1] 0.092
```

So our estimated p-value using the permutation test is 0.092 which is close to the normal theory based value of 0.0791. We could reduce variability in the estimation of the p-value simply by computing more random permutations. Since the permutation test does not depend on the assumption of normality, we might regard it as superior to the normal theory based value.

Thus it is possible to give some meaning to the p-value when the sample is the population or for samples of convenience although one has to be clear that one's conclusion apply only the particular sample.

Tests involving just one predictor also fall within the permutation test framework. We permute that predictor rather than the response

Another approach that gives meaning to the p-value when the sample is the population involves the imaginative concept of "alternative worlds" where the sample/population at hand is supposed to have been randomly selected from parallel universes. This argument is definitely more tenuous.

2. A model is usually only an approximation of underlying reality which makes the meaning of the parameters debatable at the very least. We will say more on the interpretation of parameter estimates later but the precision of the statement that $\beta_1 = 0$ exactly is at odds with the acknowledged approximate nature of the model. Furthermore, it is highly unlikely that a predictor that one has taken the trouble to measure and analyze has exactly zero effect on the response. It may be small but it won't be zero.

   This means that in many cases, we know that the point null hypothesis is false without even looking at the data. Furthermore, we know that the more data we have, the greater the power of our tests. Even small differences from zero will be detected with a large sample. Now if we fail to reject the null hypothesis, we might simply conclude that we didn't have enough data to get a significant result. According to this view, the hypothesis test just becomes a test of sample size. For this reason, I prefer confidence intervals.

3. The inference depends on the correctness of the model we use. We can partially check the assumptions about the model but there will always be some element of doubt. Sometimes the data may suggest more than one possible model which may lead to contradictory results.

4. Statistical significance is not equivalent to practical significance. The larger the sample, the smaller your p-values will be so don't confuse p-values with a big predictor effect. With large datasets it will

be very easy to get statistically significant results, but the actual effects may be unimportant. Would we really care if test scores were 0.1% higher in one state than another? Or that some medication reduced pain by 2%? Confidence intervals on the parameter estimates are a better way of assessing the size of an effect. There are useful even when the null hypothesis is not rejected because they tell us how confident we are that the true effect or value is close to the null.

Even so, hypothesis tests do have some value, not least because they impose a check on unreasonable conclusions which the data simply does not support.

3.4 Confidence Intervals for β

Confidence intervals provide an alternative way of expressing the uncertainty in our estimates. Even so, they are closely linked to the tests that we have already constructed. For the confidence intervals and regions that we will consider here, the following relationship holds. For a $100(1-\alpha)\%$ confidence region, any point that lies within the region represents a null hypothesis that would not be rejected at the $100\alpha\%$ level while every point outside represents a null hypothesis that would be rejected. So, in a sense, the confidence region provides a lot more information than a single hypothesis test in that it tells us the outcome of a whole range of hypotheses about the parameter values. Of course, by selecting the particular level of confidence for the region, we can only make tests at that level and we cannot determine the p-value for any given test simply from the region. However, since it is dangerous to read too much into the relative size of p-values (as far as how much evidence they provide against the null), this loss is not particularly important.

The confidence region tells us about plausible values for the parameters in a way that the hypothesis test cannot. This makes it more valuable.

As with testing, we must decide whether to form confidence regions for parameters individually or simultaneously. Simultaneous regions are preferable but for more than two dimensions they are difficult to display and so there is still some value in computing the one-dimensional confidence intervals.

We start with the simultaneous regions. Some results from multivariate analysis show that

$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{\sigma^2} \sim \chi^2_p$$

and

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

and these two quantities are independent. Hence

$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{p\hat{\sigma}^2} \sim \frac{\chi^2_p/p}{\chi^2_{n-p}/(n-p)} \equiv F_{p,n-p}$$

So to form a $100(1-\alpha)$ % confidence region for β, take β such that

$$(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta) \leq p\hat{\sigma}^2 F^{(\alpha)}_{p,n-p}$$

These regions are ellipsoidally shaped. Because these ellipsoids live in higher dimensions, they cannot easily be visualized.

Alternatively, one could consider each parameter individually which leads to confidence intervals which take the general form of

$$\text{estimate} \pm \text{critical value} \times \text{s.e. of estimate}$$

or specifically in this case:

$$\hat{\beta}_i \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{(X^T X)^{-1}_{ii}}$$

It's better to consider the joint confidence intervals when possible, especially when the $\hat{\beta}$ are heavily correlated.

Consider the full model for the savings data. The `.` in the model formula stands for "every other variable in the data frame" which is a useful abbreviation.

```
> g <- lm(sr ~ ., savings)
> summary(g)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.566087   7.354516    3.88  0.00033
pop15       -0.461193   0.144642   -3.19  0.00260
pop75       -1.691498   1.083599   -1.56  0.12553
dpi         -0.000337   0.000931   -0.36  0.71917
ddpi         0.409695   0.196197    2.09  0.04247

Residual standard error: 3.8 on 45 degrees of freedom
Multiple R-Squared: 0.338,      Adjusted R-squared: 0.28
F-statistic: 5.76 on 4 and 45 degrees of freedom,       p-value: 0.00079
```

We can construct individual 95% confidence intervals for the regression parameters of `pop75`:

```
> qt(0.975,45)
[1] 2.0141
> c(-1.69-2.01*1.08,-1.69+2.01*1.08)
[1] -3.8608  0.4808
```

and similarly for growth

```
> c(0.41-2.01*0.196,0.41+2.01*0.196)
[1] 0.01604 0.80396
```

Notice that this confidence interval is pretty wide in the sense that the upper limit is about 50 times larger than the lower limit. This means that we are not really that confident about what the exact effect of growth on savings really is.

Confidence intervals often have a duality with two-sided hypothesis tests. A 95% confidence interval contains all the null hypotheses that would not be rejected at the 5% level. Thus the interval for `pop75` contains zero which indicates that the null hypothesis $H_0 : \beta_{pop75} = 0$ would not be rejected at the 5% level. We can see from the output above that the p-value is 12.5% — greater than 5% — confirming this point. In contrast, we see that the interval for `ddpi` does not contain zero and so the null hypothesis is rejected for its regression parameter.

Now we construct the joint 95% confidence region for these parameters. First we load in a "library" for drawing confidence ellipses which is not part of base R:

```
> library(ellipse)
```

and now the plot:

```
> plot(ellipse(g,c(2,3)),type="l",xlim=c(-1,0))
```

add the origin and the point of the estimates:

```
> points(0,0)
> points(g$coef[2],g$coef[3],pch=18)
```

How does the position of the origin relate to a test for removing `pop75` and `pop15`?

Now we mark the one way confidence intervals on the plot for reference:

```
> abline(v=c(-0.461-2.01*0.145,-0.461+2.01*0.145),lty=2)
> abline(h=c(-1.69-2.01*1.08,-1.69+2.01*1.08),lty=2)
```

See the plot in Figure 3.3.

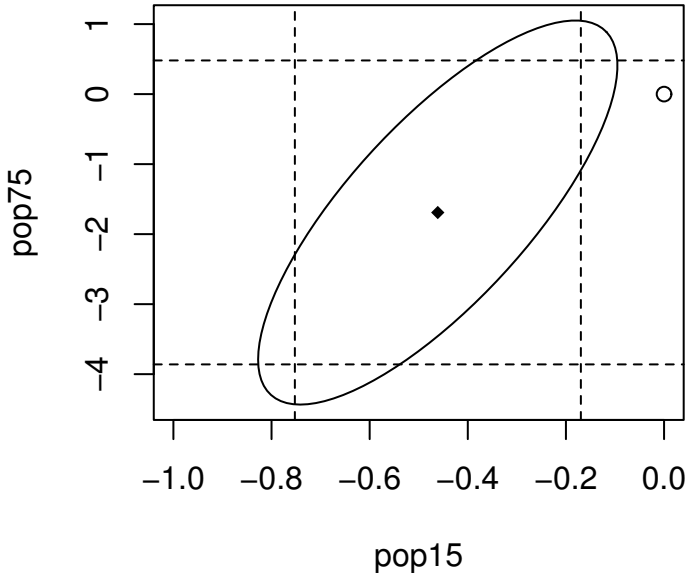

Figure 3.3: Confidence ellipse and regions for β_{pop75} and β_{pop15}

Why are these lines not tangential to the ellipse? The reason for this is that the confidence intervals are calculated individually. If we wanted a 95% chance that both intervals contain their true values, then the lines would be tangential.

In some circumstances, the origin could lie within both one-way confidence intervals, but lie outside the ellipse. In this case, both one-at-a-time tests would not reject the null whereas the joint test would. The latter test would be preferred. It's also possible for the origin to lie outside the rectangle but inside the ellipse. In this case, the joint test would not reject the null whereas both one-at-a-time tests would reject. Again we prefer the joint test result.

Examine the correlation of the two predictors:

```
> cor(savings$pop15,savings$pop75)
[1] -0.90848
```

But from the plot, we see that coefficients have a positive correlation. The correlation between predictors and the correlation between the coefficients of those predictors are often different in sign. Intuitively, this

can be explained by realizing that two negatively correlated predictors are attempting to the perform the same job. The more work one does, the less the other can do and hence the positive correlation in the coefficients.

3.5 Confidence intervals for predictions

Given a new set of predictors, x_0 what is the predicted response? Easy — just $\hat{y}_0 = x_0^T\hat{\beta}$. However, we need to distinguish between predictions of the future mean response and predictions of future observations. To make the distinction, suppose we have built a regression model that predicts the selling price of homes in a given area that is based on predictors like the number of bedrooms, closeness to a major highway etc. There are two kinds of predictions that can be made for a given x_0.

1. Suppose a new house comes on the market with characteristics x_0. Its selling price will be $x_0^T\beta + \varepsilon$. Since $E\varepsilon = 0$, the predicted price is $x_0^T\hat{\beta}$ but in assessing the variance of this prediction, we must include the variance of ε.

2. Suppose we ask the question — "What would the house with characteristics x_0" sell for on average. This selling price is $x_0^T\beta$ and is again predicted by $x_0^T\hat{\beta}$ but now only the variance in $\hat{\beta}$ needs to be taken into account.

Most times, we will want the first case which is called "prediction of a future value" while the second case, called "prediction of the mean response" is less common.

Now var $(x_0^T\hat{\beta}) = x_0^T(X^TX)^{-1}x_0\sigma^2$.

A future observation is predicted to be $x_0^T\hat{\beta} + \varepsilon$ (where we don't what the future ε will turn out to be). So a $100(1-\alpha)$ % confidence interval for a single future response is

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)}\hat{\sigma}\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$$

If on the other hand, you want a confidence interval for the average of the responses for given x_0 then use

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)}\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}$$

We return to the Galapagos data for this example.

```
> g <- lm(Species ~ Area+Elevation+Nearest+Scruz+Adjacent,data=gala)
```

Suppose we want to predict the number of species (of tortoise) on an island with predictors 0.08,93,6.0,12.0,0.34 (same order as in the dataset). Of course it is difficult to see why in practice we would want to do this because a new island is unlikely to present itself. For a dataset like this interest would center on the structure of the model and relative importance of the predictors, so we should regard this more as a "what if?" exercise.

Do it first directly from the formula:

```
> x0 <- c(1,0.08,93,6.0,12.0,0.34)
> y0 <- sum(x0*g$coef)
> y0
[1] 33.92
```

This is the predicted no. of species which is not a whole number as the response is. We could round up to 34.

Now if we want a 95% confidence interval for the prediction, we must decide whether we are predicting the number of species on one new island or the mean response for all islands with same predictors x0. Possibly, an island might not have been surveyed for the original dataset in which case the former interval would be the one we want. For this dataset, the latter interval would be more valuable for "what if?" type calculations.

First we need the t-critical value:

```
> qt(0.975,24)
[1] 2.0639
```

You may need to recalculate the $(X^TX)^{-1}$ matrix:

```
> x <- cbind(1,gala[,3:7])
> x <- as.matrix(x)
> xtxi <- solve(t(x) %*% x)
```

The width of the bands for mean response CI is

```
> bm <- sqrt(x0 %*% xtxi %*% x0) *2.064 * 60.98
> bm
      [,1]
[1,] 32.89
```

and the interval is

```
> c(y0-bm,y0+bm)
[1]  1.0296 66.8097
```

Now we compute the prediction interval for the single future response.

```
> bm <- sqrt(1+x0 %*% xtxi %*% x0) *2.064 * 60.98
> c(y0-bm,y0+bm)
[1]  -96.17  164.01
```

What physically unreasonable feature do you notice about it? In such instances, impossible values in the confidence interval can be avoided by transforming the response, say taking logs, (explained in a later chapter) or by using a probability model more appropriate to the response. The normal distribution is supported on the whole real line and so negative values are always possible. A better choice for this example might be the Poisson distribution which is supported on the non-negative integers.

There is a more direct method for computing the CI. The function `predict()` requires that its second argument be a data frame with variables named in the same way as the original dataset:

```
> predict(g,data.frame(Area=0.08,Elevation=93,Nearest=6.0,Scruz=12,
  Adjacent=0.34),se=T)
$fit:
 33.92
```

```
$se.fit:
 15.934

$df:
[1] 24

$residual.scale:
[1] 60.975
```

The width of the mean response interval can then be calculated by multiplying the se for the fit by the appropriate t-critical value:

```
> 15.934*2.064
[1] 32.888
```

which matches what we did before. CI's for the single future response could also be derived.

3.6 Orthogonality

Suppose we can partition X in two, $X = [X_1|X_2]$ such that $X_1^T X_2 = 0$. So now

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

and

$$X^T X = \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix} = \begin{pmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{pmatrix}$$

which means

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y \qquad \hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T y$$

Notice that $\hat{\beta}_1$ will be the same regardless of whether X_2 is in the model or not (and vice versa). Now if we wish to test $H_0 : \beta_1 = 0$, it should be noted that $\mathrm{RSS}_\Omega/df = \hat{\sigma}^2_\Omega$ will be different depending on whether X_2 is included in the model or not but the difference in F is not liable to be so large as in non-orthogonal cases.

Orthogonality is a desirable property but will only occur when X is chosen by the experimenter (it is a feature of a good design). In observational data, we do not have direct control over X which is the source of much of the interpretational difficulties associated with non-experimental data.

Here's an example of an experiment to determine the effects of column temperature, gas/liquid ratio and packing height in reducing unpleasant odor of chemical product that was being sold for household use.

Read the data in and display.

```
> data(odor)
> odor
   odor temp gas pack
 1   66   -1  -1    0
 2   39    1  -1    0
 3   43   -1   1    0
 4   49    1   1    0
 5   58   -1   0   -1
```

```
 6   17    1   0   -1
 7   -5   -1   0    1
 8  -40    1   0    1
 9   65    0  -1   -1
10    7    0   1   -1
11   43    0  -1    1
12  -22    0   1    1
13  -31    0   0    0
14  -35    0   0    0
15  -26    0   0    0
```

The three predictors have been transformed from their original scale of measurement, for example temp = (Fahrenheit-80)/40 so the original values of the predictor were 40,80 and 120. I don't know the scale of measurement for odor.

Here's the X-matrix:

```
> x <- as.matrix(cbind(1,odor[,-1]))
```

and X^TX:

```
> t(x) %*% x
      1 temp gas pack
   1 15    0   0    0
temp  0    8   0    0
 gas  0    0   8    0
pack  0    0   0    8
```

The matrix is diagonal. What would happen if `temp` was measured in the original Fahrenheit scale? The matrix would still be diagonal but the entry corresponding to `temp` would change.

Now fit a model:

```
> g <- lm(odor ~ temp + gas + pack, data=odor)
> summary(g,cor=T)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     15.2        9.3    1.63     0.13
temp           -12.1       12.7   -0.95     0.36
gas            -17.0       12.7   -1.34     0.21
pack           -21.4       12.7   -1.68     0.12

Residual standard error: 36 on 11 degrees of freedom
Multiple R-Squared: 0.334,      Adjusted R-squared: 0.152
F-statistic: 1.84 on 3 and 11 degrees of freedom,       p-value: 0.199

Correlation of Coefficients:
     (Intercept)     temp gas
temp   -1.52e-17
gas    -1.52e-17 4.38e-17
pack    0.00e+00 0.00e+00   0
```

Check out the correlation of the coefficients - why did that happen?. Notice that the standard errors for the coefficients are equal due to the balanced design. Now drop one of the variables:

```
> g <- lm(odor ~ gas + pack, data=odor)
> summary(g)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.20       9.26    1.64     0.13
gas           -17.00      12.68   -1.34     0.20
pack          -21.37      12.68   -1.69     0.12

Residual standard error: 35.9 on 12 degrees of freedom
Multiple R-Squared: 0.279,      Adjusted R-squared: 0.159
F-statistic: 2.32 on 2 and 12 degrees of freedom,       p-value: 0.141
```

Which things changed - which stayed the same? The coefficients themselves do not change but the residual standard error does change slightly which causes small changes in the standard errors of the coefficients, t-statistics and p-values, but nowhere near enough to change our qualitative conclusions.

That was data from an experiment so it was possible to control the values of the predictors to ensure orthogonality. Now consider the savings data which is observational:

```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> summary(g)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.566087   7.354516    3.88  0.00033
pop15       -0.461193   0.144642   -3.19  0.00260
pop75       -1.691498   1.083599   -1.56  0.12553
dpi         -0.000337   0.000931   -0.36  0.71917
ddpi         0.409695   0.196197    2.09  0.04247

Residual standard error: 3.8 on 45 degrees of freedom
Multiple R-Squared: 0.338,      Adjusted R-squared: 0.28
F-statistic: 5.76 on 4 and 45 degrees of freedom,       p-value: 0.00079
```

Drop `pop15` from the model:

```
> g <- update(g, . ~ . - pop15)
> summary(g)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.487494   1.427662    3.84  0.00037
pop75       0.952857   0.763746    1.25  0.21849
dpi         0.000197   0.001003    0.20  0.84499
ddpi        0.473795   0.213727    2.22  0.03162
```

```
Residual standard error: 4.16 on 46 degrees of freedom
Multiple R-Squared: 0.189,      Adjusted R-squared: 0.136
F-statistic: 3.57 on 3 and 46 degrees of freedom,       p-value: 0.0209
```

What changed? By how much? Pay particular attention to `pop75`. The effect has now become positive whereas it was negative. Granted, in neither case is it significant, but it is not uncommon in other datasets for such sign changes to occur and for them to be significant.

3.7 Identifiability

The least squares estimate is the solution to the normal equations:

$$X^T X\hat{\beta} = X^T y$$

where X is an $n \times p$ matrix. If $X^T X$ is singular and cannot be inverted, then there will be infinitely many solutions to the normal equations and $\hat{\beta}$ is at least partially unidentifiable.

Unidentifiability will occur when X is not of full rank — when its columns are linearly dependent. With observational data, unidentifiability is usually caused by some oversight: Here are some examples:

1. A person's weight is measured both in pounds and kilos and both variables are entered into the model.

2. For each individual we record no. of years of education K-12 and no. of years of post-HS education and also the total no. of years of education and put all three variables into the model.

3. $p > n$ — more variables than cases. When $p = n$, we may perhaps estimate all the parameters, but with no degrees of freedom left to estimate any standard errors or do any testing. Such a model is called *saturated*. When $p > n$, then the model is called *supersaturated*. Oddly enough, such models are considered in large scale screening experiments used in product design and manufacture, but there is no hope of uniquely estimating all the parameters in such a model.

Such problems can be avoided by paying attention. Identifiability is more of an issue in designed experiments. Consider a simple two sample experiment:

	Response
Treatment	$y_1, \ldots, y_n$
Control	$y_{n+1}, \ldots, y_{m+n}$

Suppose we try to model the response by an overall mean μ and group effects α_1 and α_2:

$$y_j = \mu + \alpha_i + \varepsilon_j \qquad i = 1,2 \quad j = 1, \ldots, m+n$$

$$\begin{pmatrix} y_1 \\ \cdots \\ y_n \\ y_{n+1} \\ \cdots \\ y_{m+n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ & \cdots & \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ . & . & . \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \cdots \\ \cdots \\ \cdots \\ \varepsilon_{m+n} \end{pmatrix}$$

Now although X has 3 columns, it has only rank 2 — $(\mu, \alpha_1, \alpha_2)$ are not identifiable and the normal equations have infinitely many solutions. We can solve this problem by imposing some constraints, $\mu = 0$ or $\alpha_1 + \alpha_2 = 0$ for example.

Statistics packages handle non-identifiability differently. In the regression case above, some may return error messages and some may fit models because rounding error may remove the exact identifiability. In other cases, constraints may be applied but these may be different from what you expect.

Identifiability means that

1. You have insufficient data to estimate the parameters of interest *or*
2. You have more parameters than are necessary to model the data.

Here's an example. Suppose we create a new variable for the savings dataset - the percentage of people between 15 and 75:

```
> pa <- 100-savings$pop15-savings$pop75
```

and add that to the model:

```
> g <- lm(sr ~ pa + pop15 + pop75 + dpi + ddpi, data=savings)
> summary(g)
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.41e+02   1.03e+02   -1.37    0.177
pa           1.69e+00   1.08e+00    1.56    0.126
pop15        1.23e+00   9.77e-01    1.26    0.215
dpi         -3.37e-04   9.31e-04   -0.36    0.719
ddpi         4.10e-01   1.96e-01    2.09    0.042
```

We get a message about one undefined coefficient because the rank of the design matrix X is 5 but should be 6.

Let's take a look at the X-matrix:

```
> x <- as.matrix(cbind(1,pa,savings[,-1]))
> dimnames(x) <- list(row.names(savings),c("int","pa","p15","p75",
  "dpi","ddpi"))
```

If we didn't know which linear combination was causing the trouble, how would we find out? An eigen decomposition of X^TX can help:

```
> e <- eigen(t(x) %*% x)
> signif(e$values,3)
[1] 1.10e+08 1.10e+05 3.19e+03 3.74e+02 1.37e+01 1.09e-14
> signif(e$vectors,3)
int  0.000506  0.0141 -0.00125  0.000603  0.00989  1.00e+00
pa   0.034300  0.7940  0.59700  0.098100 -0.05630 -1.00e-02
p15  0.014700  0.6040 -0.79500 -0.031000  0.04800 -1.00e-02
p75  0.001610  0.0164  0.07310 -0.006840  0.99700 -1.00e-02
dpi  0.999000 -0.0363 -0.00906 -0.001170 -0.00036 -1.07e-17
ddpi 0.001740  0.0594  0.08310 -0.995000 -0.01390 -4.97e-16
```

Only the last eigenvalue is zero, indicating one linear combination is the problem. We can determine which linear combination from the last eigenvalue (last column of the matrix. From this we see that `100-pa-p15-p75=0` is the offending combination.

Lack of identifiability is obviously a problem but it is usually easy to identify and work around. More problematic are cases where we are close to unidentifiability. To demonstrate this, suppose we add a small random perturbation to the third decimal place of `pa` by adding a random variate from $U[-0.005, 0.005]$ where U denotes the uniform distribution:

```
> pae <- pa +0.001*(runif(50)-0.5)
```

and now refit the model:

```
> ge <- lm(sr ~ pae+pop15+pop75+dpi+ddpi,savings)
> summary(ge)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.57e+05   1.81e+05    0.87    0.391
pae         -1.57e+03   1.81e+03   -0.87    0.391
pop15       -1.57e+03   1.81e+03   -0.87    0.391
pop75       -1.57e+03   1.81e+03   -0.87    0.390
dpi         -3.34e-04   9.34e-04   -0.36    0.722
ddpi         4.11e-01   1.97e-01    2.09    0.042
```

Notice the now all parameters can be estimated but the standard errors are very large because we cannot estimate them in a stable way. We deliberately caused this problem so we know the cause but in general we need to be able to identify such situations. We do this in Chapter 9.

3.8 Summary

We have described a linear model $y = X\beta + \varepsilon$. The parameters β may be estimated using least squares $\hat{\beta} = (X^T X)^{-1} X^T y$. If we further assume that $\varepsilon \sim N(0, \sigma^2 I)$ then we can test any linear hypothesis about β, construct confidence regions for β, make predictions with confidence intervals.

3.9 What can go wrong?

Many things, unfortunately — we try to categorize them below:

3.9.1 Source and quality of the data

How the data was collected directly effects what conclusions we can draw.

1. We may have a biased sample, such as a sample of convenience, from the population of interest. This makes it very difficult to extrapolate from what we see in the sample to general statements about the population. As we have seen, in some cases the sample is the population, in which case any generalization of the conclusions is problematic.

2. Important predictors may not have been observed. This means that our predictions may be poor or we may misinterpret the relationship between the predictors and the response.

3. Observational data make causal conclusions problematic — lack of orthogonality makes disentangling effects difficult. Missing predictors add to this problem.

4. The range and qualitative nature of the data may limit effective predictions. It is unsafe to extrapolate too much. Carcinogen trials may apply large doses to mice. What do the results say about small doses applied to humans? Much of the evidence for harm from substances such as asbestos and radon comes from people exposed to much larger amounts than that encountered in a normal life. It's clear that workers in older asbestos manufacturing plants and uranium miners suffered from their respective exposures to these substances, but what does that say about the danger to you or I?

3.9.2 Error component

We hope that $\varepsilon \sim N(0, \sigma^2 I)$ but

1. Errors may be heterogeneous (unequal variance).

2. Errors may be correlated.

3. Errors may not be normally distributed.

The last defect is less serious than the first two because even if the errors are not normal, the $\hat{\beta}$'s will tend to normality due to the power of the central limit theorem. With larger datasets, normality of the data is not much of a problem.

3.9.3 Structural Component

The structural part of linear model, $Ey = X\beta$ may be incorrect. The model we use may come from different sources:

1. Physical theory may suggest a model, for example Hooke's law says that the extension of a spring is proportional to the weight attached. Models like these usually arise in the physical sciences and engineering.

2. Experience with past data. Similar data used in the past was modeled in a particular way. It's natural to see if the same model will work the current data. Models like these usually arise in the social sciences.

3. No prior idea - the model comes from an exploration of the data itself.

Confidence in the conclusions from a model declines as we progress through these. Models that derive directly from physical theory are relatively uncommon so that usually the linear model can only be regarded as an approximation to a reality which is very complex.

Most statistical theory rests on the assumption that the model is correct. In practice, the best one can hope for is that the model is a fair representation of reality. A model can be no more than a good portrait.

> All models are wrong but some are useful. *George Box*

is only a slight exaggeration. Einstein said

> So far as theories of mathematics are about reality; they are not certain; so far as they are certain, they are not about reality.

3.10 Interpreting Parameter Estimates

Suppose we fit a model to obtain the regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

What does $\hat{\beta}_1$ mean? In some case, a β might represent a real physical constant, but often the statistical model is just a convenience for representing a complex reality and so the real meaning of a particular β is not obvious.

Let's start with a naive interpretation: "A unit change in x_1 will produce a change of $\hat{\beta}_1$ in the response".

For a properly designed experiment, this interpretation is reasonable provided one pays attention to concerns such as extrapolation and appropriateness of the model selected. The effects of other variables that are included in the experiment can separated out if an orthogonal design is used. For variables not included in the experiment by choice, we may eliminate their effect by holding them constant. If variables that impact the response are not included because they are not known, we use randomization to control their effect. The treatments (predictor values) are assigned to the experimental units or subjects at random. This ensures that these unknown variables will not be correlated in expectation with the predictors we do examine and allows us to come to causal conclusions. These unknown predictors do not, on the average, affect the parameter estimates of interest, but they do contribute to the residual standard error so it's sometimes better to incorporate them in the experimental design if the become known, as this allows for more precise inference.

In a few tightly controlled experiments, it is possible to claim that measurement error is the only kind of error but usually some of the "error" actually comes from the effects of unmeasured variables. We can decompose the usual model as follows:

$$\begin{aligned} y &= X\beta + \varepsilon \\ &= X\beta + Z\gamma + \delta \end{aligned}$$

where Z are unincluded predictors and δ is measurement error in the response. We can assume that $E\varepsilon = 0$ without any loss of generality, because if $E\varepsilon = c$, we could simply redefine β_0 as $\beta_0 + c$ and the error would again have expectation zero. This is another reason why it is generally unwise to remove the intercept term from the model since it acts as a sink for the mean effect of unincluded variables. So we see that ε incorporates both measurement error and the effect of other variables. In a designed experiment, provided the assignment of the experimental units is random, we have $cor(X,Z) = 0$ so that the estimate of β is unaffected in expectation by the presence of Z.

For observational data, no randomization can be used in assigning treatments to the units and orthogonality won't just happen. There are serious objections to any causal conclusions. An inference of causality is often desired but this is usually too much to expect from observational data. An unmeasured and possible unsuspected "lurking" variable Z may be the real cause of an observed relationship between y and X. See Figure 3.4. For example, we will observe a positive correlation among the shoe sizes and reading abilities of elementary school students but this relationship is driven by a lurking variable — the age of the child.

So in observational studies, because we have no control over the assignment of units, we have $cor(X,Z) \neq 0$ and the observed or worse, unobserved, presence of Z causes us great difficulty. In Figure 3.5, we see the effect of possible confounding variables demonstrated.

In observational studies, it is important to adjust for the effects of possible confounding variables such as the Z shown in Figure 3.5. If such variables can be identified, then at least their effect can be interpreted. Unfortunately, one can never be sure that the all relevant Z have been identified.

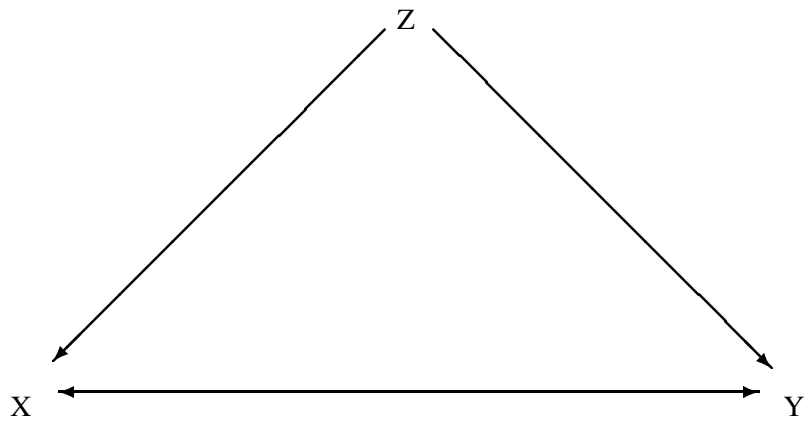

Figure 3.4: Is the relationship between x and y, really caused by z?

What if all relevant variables have been measured? In other words, suppose there are no unidentified *lurking* variables. Even then the naive interpretation does not work. Consider

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

but suppose we change $x_2 \rightarrow x_1 + x_2$ then

$$y = \hat{\beta}_0 + (\hat{\beta}_1 - \hat{\beta}_2) x_1 + \hat{\beta}_2 (x_1 + x_2)$$

The coefficient for x_1 has changed. Interpretation cannot be done separately for each variable. This is a practical problem because it is not unusual for the predictor of interest, x_1 in this example, to be mixed up in some way with other variables like x_2.

Let's try a new interpretation:

"$\hat{\beta}_1$ is the effect of x_1 when all the other (specified) predictors are held constant".

This too has problems. Often in practice, individual variables cannot be changed without changing others too. For example, in economics we can't expect to change tax rates without other things changing too. Furthermore, this interpretation requires the specification of the other variables - changing which other variables are included will change the interpretation. Unfortunately, there is no simple solution.

Just to amplify this consider the effect of `pop75` on the savings rate in the savings dataset. I'll fit four different models, all including `pop75` but varying the inclusion of other variables.

```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
> summary(g)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.566087   7.354516    3.88  0.00033
pop15       -0.461193   0.144642   -3.19  0.00260
pop75       -1.691498   1.083599   -1.56  0.12553
dpi         -0.000337   0.000931   -0.36  0.71917
ddpi         0.409695   0.196197    2.09  0.04247

Residual standard error: 3.8 on 45 degrees of freedom
Multiple R-Squared: 0.338,      Adjusted R-squared: 0.28
F-statistic: 5.76 on 4 and 45 degrees of freedom,       p-value: 0.00079
```

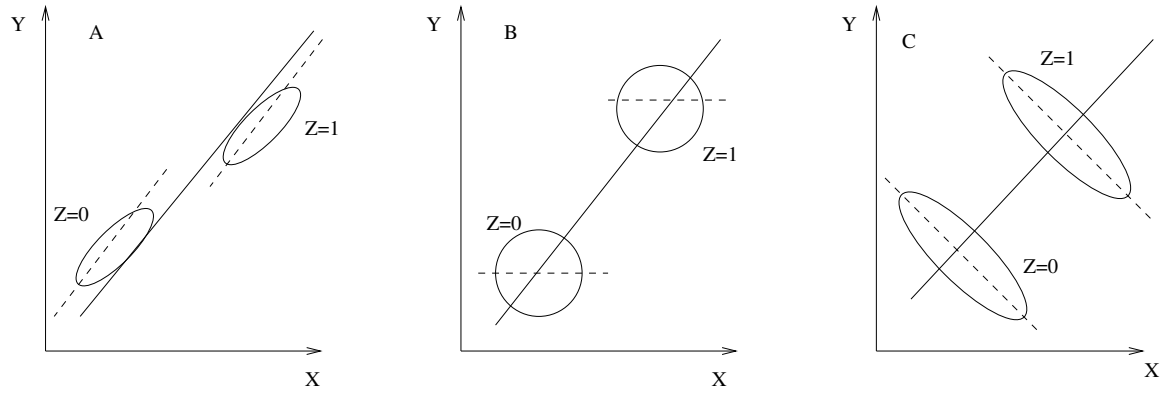

Figure 3.5: Possible confounding effects illustrated. Imagine the data is observed within the ellipses. If the effect of Z is ignored, a strong positive correlation between X and Y is observed in all three cases. In panel A, we see that when we allow for the effect of Z by observing the relationship between X and Y separately within each level of Z, that the relationship remains a positive correlation. In panel B, after allowing for Z, there is no correlation between X and Y, while in panel C, after allowing for Z, the relationship becomes a negative correlation.

It is perhaps surprising that `pop75` is not significant in this model. However, `pop75` is negatively correlated with `pop15` since countries with proportionately more younger people are likely to relatively fewer older ones and vice versa. These two variables are both measuring the nature of the age distribution in a country. When two variables that represent roughly the same thing are included in a regression equation, it is not unusual for one (or even both) of them to appear insignificant even though prior knowledge about the effects of these variables might lead one to expect them to be important.

```
> g2 <- lm(sr ~ pop75 + dpi + ddpi, data=savings)
> summary(g2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.487494   1.427662    3.84  0.00037
pop75       0.952857   0.763746    1.25  0.21849
dpi         0.000197   0.001003    0.20  0.84499
ddpi        0.473795   0.213727    2.22  0.03162

Residual standard error: 4.16 on 46 degrees of freedom
Multiple R-Squared: 0.189,      Adjusted R-squared: 0.136
F-statistic: 3.57 on 3 and 46 degrees of freedom,       p-value: 0.0209
```

We note that the income variable `dpi` and `pop75` are both not significant in this model and yet one might expect both of them to have something to do with savings rates. Higher values of these variables are both associated with wealthier countries. Let's see what happens when we drop `dpi` from the model:

```
> g3 <- lm(sr ~ pop75 + ddpi, data=savings)
> summary(g3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.470      1.410    3.88  0.00033
```

```
pop75          1.073      0.456    2.35  0.02299
ddpi           0.464      0.205    2.26  0.02856

Residual standard error: 4.12 on 47 degrees of freedom
Multiple R-Squared: 0.188,      Adjusted R-squared: 0.154
F-statistic: 5.45 on 2 and 47 degrees of freedom,       p-value: 0.00742
```

Now `pop75` is statistically significant with a positive coefficient. We try dropping `ddpi`:

```
> g4 <- lm(sr ~ pop75, data=savings)
> summary(g4)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.152      1.248    5.73  6.4e-07
pop75          1.099      0.475    2.31    0.025

Residual standard error: 4.29 on 48 degrees of freedom
Multiple R-Squared:  0.1,       Adjusted R-squared: 0.0814
F-statistic: 5.34 on 1 and 48 degrees of freedom,       p-value: 0.0251
```

The coefficient and p-value do not change much here due to the low correlation between `pop75` and `ddpi`.

Compare the coefficients and p-values for `pop75` throughout. Notice how the sign and significance change in Table3.2.

No. of Preds	Sign	Significant?
4	-	no
3	+	no
2	+	yes
1	+	yes

Table 3.2: Sign and Significance of $\hat{\beta}_{pop75}$

We see that the significance and the direction of the effect of `pop75` change according to what other variables are also included in the model. We see that no simple conclusion about the effect of `pop75` is possible. We must find interpretations for a variety of models. We certainly won't be able to make any causal conclusions.

In observational studies, there are steps one can take to make a stronger case for causality:

1. Try to include all relevant variables
2. Use non-statistical knowledge of the physical nature of the relationship.
3. Try a variety of models - see if a similar effect is observed. Is $\hat{\beta}_1$ similar, no matter what the model?
4. Multiple studies under different conditions can help confirm a relationship. The connection between smoking and lung cancer was suspected since the early 50's but other explanations for the effect were proposed. It was many years before other plausible explanations were eliminated.

   The news media often jump on the results of a single study but one should be suspicious of these one off results. Publication bias is a problem. Many scientific journal will not publish the results of a

study whose conclusions do not reject the null hypothesis. If different researchers keep studying the same relationship, sooner or later one of them will come up with a significant effect even if one really doesn't exist. It's not easy to find out about all the studies with negative results so it is easy to make the wrong conclusions.

Another source of bias is that researchers have a vested interest in obtaining a positive result. There is often more than one way to analyze the data and the researchers may be tempted to pick the one that gives them the results they want. This is not overtly dishonest but it does lead to a bias towards positive results.

It's difficult to assess the evidence in these situations and one can never be certain. The history of the study of the link between smoking and lung cancer shows that it takes a great deal of effort to progress beyond the observation of an association to strong evidence of causation. One can never be 100% sure.

An alternative approach is recognize that the parameters and their estimates are fictional quantities in most regression situations. The "true" values may never be known (if they even exist in the first place). Instead concentrate on predicting future values - these may actually be observed and success can then be measured in terms of how good the predictions were.

Consider a prediction made using each of the four models above:

```
> x0 <- data.frame(pop15=32,pop75=3,dpi=700,ddpi=3)
> predict(g,x0)
[1] 9.7267
> predict(g2,x0)
[1] 9.9055
> predict(g3,x0)
[1] 10.078
> predict(g4,x0)
[1] 10.448
```

Prediction is more stable than parameter estimation. This enables a rather cautious interpretation of $\hat{\beta}_1$. Suppose the predicted value of y is $\hat{y}$ for given x_1 and for other given predictor values. Now suppose we observe $x_1 + 1$ and the same other given predictor values then the predicted response is increased by $\hat{\beta}_1$. Notice that I have been careful to not to say that we have taken a specific individual and increased their x_1 by 1, rather we have observed a new individual with predictor $x_1 + 1$. To put it another way, people with yellow fingers tend to be smokers but making someone's fingers yellow won't make them more likely to smoke.

Prediction is conceptually simpler since interpretation is not an issue but you do need to worry about extrapolation.

1. Quantitative extrapolation: Is the new x_0 within the range of validity of the model. Is it close to the range of the original data? If not, the prediction may be unrealistic. Confidence intervals for predictions get wider as we move away from the data. We can compute these bands for our last model:

   ```
   > grid <- seq(0,10,0.1)
   > p <- predict(g4,data.frame(pop75=grid),se=T)
   > cv <- qt(0.975,48)
   > matplot(grid,cbind(p$fit,p$fit-cv*p$se,p$fit+cv*p$se),lty=c(1,2,2),
     type="l",xlab="pop75",ylab="Saving")
   > rug(savings$pop75)
   ```

We see that the confidence bands in Figure 3.6 become wider as we move away from the range of the data. However, this widening does not reflect the possibility that the structure of the model itself may change as we move into new territory. The uncertainty in the parametric estimates is allowed for but not uncertainty about the model itself. In Figure 3.7, we see that a model may fit well in the range of the data, but outside of that range, the predictions may be very bad.

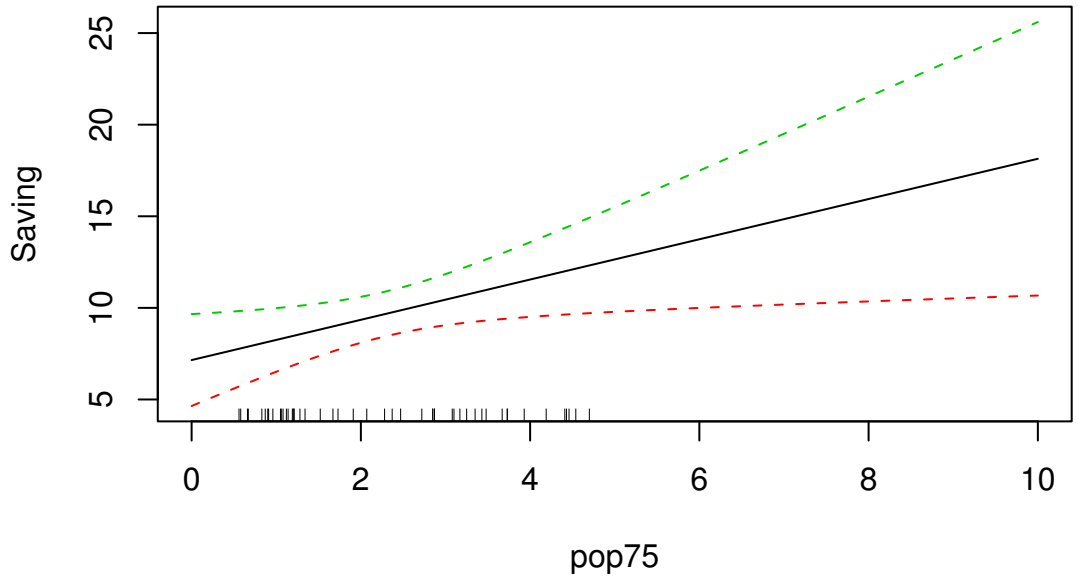

Figure 3.6: Predicted `pop75` over a range of values with 95% pointwise confidence bands for the mean response shown as dotted lines. A "rug" shows the location of the observed values of `pop75`

2. Qualitative extrapolation: Is the new x_0 drawn from the same population from which the original sample was drawn. If the model was built in the past and is to be used for future predictions, we must make a difficult judgment as to whether conditions have remained constant enough for this to work.

Let's end with a quote from the 4th century. Prediction is a tricky business — perhaps the only thing worse than a prediction is no prediction at all.

> The good Christian should beware of mathematicians and all those who make empty prophecies. The danger already exists that mathematicians have made a covenant with the devil to darken the spirit and confine man in the bonds of Hell. - St. Augustine

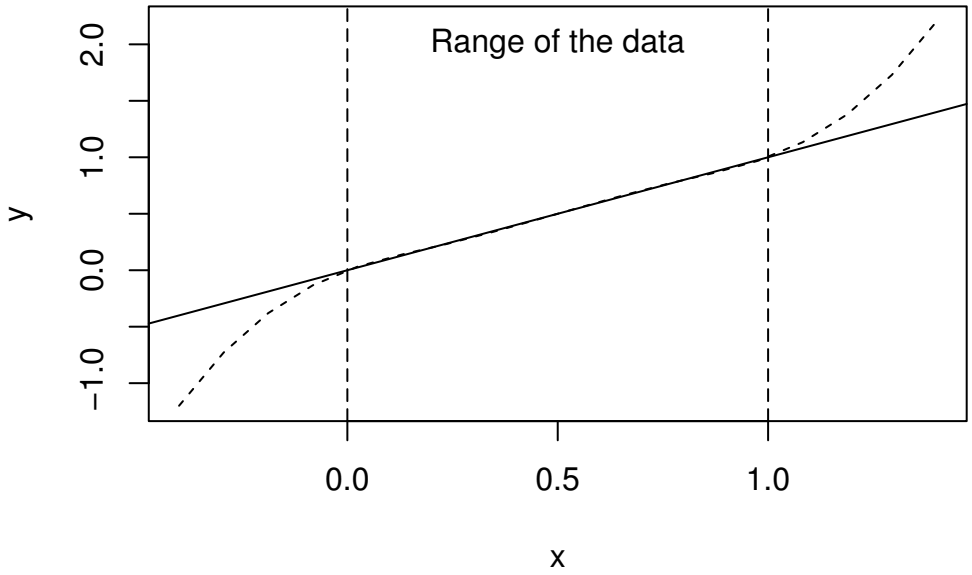

Figure 3.7: Dangers of extrapolation: The model is shown in solid, the real relationship by the dotted line. The data all lie in the predictor range [0,1]