Chapter 1

Mathematical Notation

Joseph Guinness - BTRY 6020

The language of mathematics is a system of numbers, letters, and other symbols. Like any language–written, spoken, computer–the goal of mathematical notation is clear and effective communication of ideas. In written and spoken languages, we aim to communicate complete thoughts via sentences. In computer languages, we aim to communicate to the computer what calculations should be done and printed onto the screen. In mathematical notation, we aim to communicate to one another how variables are related via equations.

In this statistics course, we are more formal about how we express our statistical models in mathematical notation. It will take some practice to master this language, but in the end you will find that the more formal approach is an effective way to avoid ambiguities in communicating your assumptions to others. And most will find it rewarding to master a new way of communication.

We will be using subscript and summation notation throughout the course. This chapter explains how this notation works. The dataset below has 5 different marathon times in minutes.

i	y_i
1	165.3
2	158.8
3	173.9
4	190.0
5	211.6

In this dataset, i is used to signify one of the integers between 1 and 5, and the subscript notation y_i represents the data value in row i of the dataset. We can think of i as a mechanism for labeling the response y. This makes it easy to refer to a specific data value, e.g. $y_3 = 173.9$.

A big part of statistical analysis is the calculation of summaries of the data, also known as statistics. We will use summation notation for sums. A simple example is

$$\sum_{i=1}^{5} y_i = y_1 + y_2 + y_3 + y_4 + y_5.$$

Underneath the Σ, we define an indexing variable i and a starting value for the indexing variable $i = 1$. Above the sum is a finishing value for the indexing variable $i = 5$. To the right of the sum is the value that is added to the sum at each specific value of i. In words, the summation notation means, "for all integers between and including $i = 1$ to $i = 5$, add up the values y_i." This is expressed explicitly on the right side of the equation.

It is important to note that there is nothing special about the fact that we picked i as the indexing variable. Even though our dataset used i to label the first column, the following is functionally the same expression:

$$\sum_{j=1}^{5} y_j = y_1 + y_2 + y_3 + y_4 + y_5.$$

We can think of the indexing variable as a temporary variable; it only has a meaning within the sum, and as long as it doesn't conflict with any of the variables that we are summing over, it doesn't matter what we pick. We could have chosen $k = 1$ to 5, or $\alpha = 1$ to 5, but not $y = 1$ to 5 because y conflicts with the the variable in the sum. y_y doesn't make any sense.

To practice, write out these sums explicitly:

$$\sum_{i=1}^{5} i =$$

$$\sum_{i=-1}^{4} i =$$

$$\sum_{i=1}^{6} (i-2) =$$

$$\sum_{i=1}^{5} i^2 =$$

$$\sum_{i=0}^{3} (-1)^i =$$

Note that in the third sum, we put parentheses around $i - 2$. This is because the following would have been ambiguous:

$$\sum_{i=1}^{6} i - 2 =$$

Without the parentheses, we don't know if this means that we added up i from 1 to 6 then subtract 2, or we add up $i-2$ from 1 to 6. These two interpretations would give different answers.

To practice, write the following explicit expressions in summation notation:

$$\begin{aligned} 1+1/2+1/3+1/4+1/5 &= \\ 1/2+1/4+1/8+1/16 &= \\ 5+4+3+2+1 &= \\ 2+4+6+8+10+12 &= \\ y_2+y_4+y_6+y_5 &= \end{aligned}$$

When we analyze more interesting datasets, our notation will have to expand to meet the increased complexity. We will use double subscripting and indexing functions. Consider the following dataset:

observation	runner	race	time
1	1	1	162.4
2	1	2	163.1
3	1	3	160.0
4	2	1	175.6
5	2	2	176.7
6	2	3	172.2

We have six observations, the first three coming from one runner, and the second three coming from another runner. Each of the two runners ran the same three races. We have a choice about the notation we use for the responses.

One choice is to label the marathon times as

$$y_1, y_2, y_3, y_4, y_5, y_6,$$

using the labels from the first column to uniquely identify the observations. In this case, y_i is the marathon time for the ith observation in the dataset, so for example, $y_2 = 163.1$ and $y_6 = 172.2$.

Another option is to use the runner and race variables to give unique identifiers to the observations. In the same order, the observations could be labeled using double subscripts as

$$y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}.$$

In this case y_{ij} is the marathon time for runner i in race j, so for example, $y_{12} = 163.1$ and $y_{23} = 172.2$. One should be careful with this notation because the same runner might run the same race twice in different years. In that case, we could have a third index to identify the year.

Either single or double subscript notation is reasonable, and we will use both in this course. The important thing is that we clearly communicate the meaning of each symbol in our notation.

We will also sometimes need a way to refer to, for example, the runner that produced the 2nd observation in the dataset. For that, we will use indexing functions. Let's set this up. Let i refer to the row of the dataset, and let $j(i)$ be a function that takes in the row as an argument, and returns the label for the runner in that row. This means that

$$j(1) = 1, \quad j(2) = 1, \quad j(3) = 1, \quad j(4) = 2, \quad j(5) = 2, \quad j(6) = 2$$

We could define a similar function $k(i)$ to identify the race where the ith observation was produced. For practice, write out $k(i)$ for $i = 1$ to $i = 6$.

To give a preview of how this is useful, could have two parameters b_1 and b_2 representing the expected time for runner 1 and runner 2. Then the residual for the ith observation is

$$e_i = y_i - b_{j(i)}.$$

In our dataset, we have for example $e_2 = y_2 - b_1$. Altenatively, if we use double indexing for the times, then the residual could be defined as

$$e_{ij} = y_{ij} - b_i.$$

Note that in the first residual expression, $j(i)$ refers to the runner that produced the ith observation, whereas in the second residual expression j is the race. observation. Either choice is acceptable notation; just make sure to define your symbols clearly.

When we have two indexing variables, we sometimes want to take sums over both of the variables. We can accomplish this with double summation. Let's start with a simple example to explain what it means.

$$\sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij} = \sum_{i=1}^{2}\Big(\sum_{j=1}^{3} y_{ij}\Big) = (y_{11} + y_{12} + y_{13}) + (y_{21} + y_{22} + y_{23})$$

When we write a double summation, we mean that for each value of the indexing variable in the first summation, we compute the second summation over the second indexing variable. To complete the sum, we add up all the results from each value of the first indexing variable. We can see how this works in the example above. In the first set of parentheses, we are adding up y_{ij} over j with i set to 1. In the second set of parentheses, we are adding up y_{ij} over j with i set to 2. The double sum is the sum of these two sums.

think about our notation y_{ij} for the time from runner i in race j. If we wanted to take the average of all performances, we would do the double summation over

both i and j and divide by the total number of observations

$$\frac{1}{6}\sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij}.$$

Chapter 2

Reasoning with Quantitative Evidence

Joseph Guinness - BTRY 6020

Statistics provides a logical framework and a set of tools that allow us to weigh the evidence contained in quantitative information. This is the primary contribution that statistics has made to human progress. Statistical analyses can be a part of, but are not a substitute for, a reasoned process of drawing conclusions from information. If you find it difficult to work through the logic behind a hypothesis test or the construction of a confidence interval, take solace in the fact that it took humans a very long time to figure this stuff out. Mathematicians mastered Fourier series, complex analysis, differential equations, and a host of other difficult problems before they figured out what the average of a sample says about the world. Part of what makes statistical analyses challenging is that most of the inferences we make from quantitative information are inductive rather than deductive; rarely are we able to say that we proved something to be true by collecting data. Most often we talk about evidence supporting one claim rather than another.

The following is a simplified workflow describing how statisticians are trained to use data to answer questions.

1. Formulate a question
2. Design a study and collect data
3. Choose a statistical model for the data
4. Use data to estimate and make inferences about model parameters
5. Make a judgment about the answer to the question based on the evidence

These tasks are all interrelated; the data collected should be relevant to the question, the model should be appropriate for the data and contain parameters (i.e. unknown components) relevant to the question, and the estimates and inferences are informed by the model assumptions and the data. In this course, we will focus mostly on the third and fourth aspects: picking models and estimating parameters, though we will work in some discussion about how to design studies that give us the best chance of answering our questions.

2.1 A Question, Study, and Decision Rule

As of 2018, the Cornell University "Facts" website (`www.cornell.edu/about/facts.cfm`) stated that 26% of Cornell students call New York State their region of origin, defined as their home at the time of matriculation. This number represents an overall percentage considering all undergraduate, graduate, and professional students. Probably if we broke the total population of students down into subpopulations, the percentage of NY students would differ. We might ask the **question** *do 26% of students in CALS call NY their region of origin?*. Since it would be difficult for us to commission a census of all CALS students, we might **design a small study** to survey a sample of students, such as the students in this class. Then we could calculate the percentage of NY students in this class, and compare that to the reported overall percentage of 26%. A first question to ask is whether the data we obtain are relevant to the question. Maybe. It depends on whether the population of students in this class is representative of the total population of CALS students, with respect to their region of origin. A surefire way to make sure a sample is representative is to select a random sample, but that is not feasible here, so we are left to grapple with our potentially non-representative sample. Already things are getting a little messy, but we can try to reason about whether the sample is representative enough, and proceed with the study, keeping in mind this potential flaw and an idea of how much it might affect the results.

Let p be the proportion of NY students in CALS. Then we can reframe our question more succinctly as, does $p = 0.26$? Suppose there are n students in the class. Our data consist of $y_1, \ldots, y_n$, where $y_i = 1$ if student i is from NY, and $y_i = 0$ otherwise. Then the proportion of NY students in this class is

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

In this course, we typically denote data with lowercase letters, although there will be exceptions to this convention. Almost certainly, $\widehat{p}$ will not exactly equal 0.26. Even if we did get exactly $\widehat{p} = 0.26$, we still would not be able to claim for sure that $p = 0.26$. At best, we can make statements about whether $p = 0.26$ is plausible or not plausible.

OK, so if we got exactly $\widehat{p} = 0.26$, the conclusion should clearly be that $p = 0.26$ is plausible. But how far away from 0.26 does $\widehat{p}$ need to be in order to claim that $p = 0.26$ is not plausible? For now, let's leave the precise answer to that question alone, and simply define the *form* that an answer will take. The form of the answer is that $p = 0.26$ is plausible if $|\widehat{p} - 0.26| < c$, and not plausible if $|\widehat{p} - 0.26| \geq c$. Statistical modeling will show its usefulness in helping us decide the threshold c.

2.2 Statistical Model

We all have some understanding of what a model is: a simplified representation of how the world works. Climate models represent the earth system with differential equations and use supercomputers to solve the equations; economic models incorporate the preferences and behaviors of various actors in an economy; species-prey models posit dynamic relationships among animal populations. None are meant to be taken literally as a statement about how the world works. Nonetheless, models are useful tools in the pursuit of understanding how the world works. But what makes a model statistical? This is not an easy question; in 2001, 86 pages in the Annals of Statistics were devoted to an article, discussion, and rejoinder trying to answer it [1]. Thus it is difficult and maybe misleading to give a one sentence definition, but we can try. A statistical model is a family of probability distributions meant to represent an assumption about how data are generated. Statistical models serve to formalize our assumptions about data-generating mechanisms using probability, and probability gives us a mathematics and a language for handling uncertainty. Like other types of models, they are not meant to be taken literally, but they serve as useful tools for evaluating quantitative evidence.

For our region of origin study, let Y_i be a random variable (RV) that takes value 1 with probability p and value 0 with probability $1 - p$, that is

$$P(Y_i = 1) = p, \quad P(Y_i = 0) = 1 - p,$$

and let the Y_i's be identically distributed, i.e. each $Y_1, \ldots, Y_n$ has the same probability distribution. This is a "named" probability distribution known as the Binomial$(1, p)$ distribution. Random variables are different from regular variables in that the "value" of a random variable is represented as a probability distribution rather than a single number, like our data y_i. We typically use uppercase letters for RVs, although again there will be exceptions. We think that Y_i is a reasonable model for y_i because they both can take on only the values 0 and 1, and the model contains a parameter p that, when estimated, can help us answer our question of interest. We say that this is family of probability distributions because each different p gives a different distribution. The statistical model has not been fully specified, however, until we say how the collection of random variables $Y_1, \ldots, Y_n$ are related to one another. This

is often a subtly tricky part. For example, students that come from the same region may be more likely to be friends, and since they're friends they may be more likely to enroll in similar courses. So if we knew that student i and student j were friends, and we knew that $Y_i = 1$, then we might believe that $P(Y_j = 1|Y_i = 1) > p$. For now, we assume that the random variables are independent. Independence has a formal mathematical definition, but here just take it to mean that knowing the realized value of Y_i does not change our beliefs about any other Y_j. The two assumptions of independence and identical distributions are abbreviated as *i.i.d.*, and so our specification of the statistical model can be stated as

$$Y_i \overset{iid}{\sim} \text{Binomial}(1, p) \quad \text{for } i = 1, \ldots, n.$$

2.3 Sampling Distributions

Once we have our statistical model, we can do calculations with it. Of particular use is to calculate the probability distribution of

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Since each Y_i is a random variable, and $\widehat{p}$ is simply a function of the Y_i's, then $\widehat{p}$ is also a random variable. Note that we have just committed a notational crime by using the same letter for $\widehat{p} = 1/n \sum y_i$ and $\widehat{p} = 1/n \sum Y_i$, which are two different things! The first is a function of the data, and simply a number, while the second is a random variable. We have to learn to live with this kind of offensive notational inconsistency because everybody else does it. It is especially bad here because students tend to struggle when thinking about the distinction between the data version and the random version. Unfortunately, the notation makes this harder for us, but luckily, we can usually tell from context whether we are referring to the data version or the random variable version. By the way, we call the data version the estimate, and we call the random variable version the estimator. At least there are two different words for them.

Since the estimator $\widehat{p}$ is a random variable, we can calculate its probability distribution. It's actually easier to calculate the distribution of $S = \sum Y_i$; we'll divide by n later if we need to. The first thing to figure out is what values S can take. It has to be an integer, since it is the sum of integers, and it has to fall between 0 (in the case that no students are from NY) and n (in the case that all students are from NY). Because of the *i.i.d.* assumption, S has a Binomial(n, p) distribution, which has probability mass function (pmf)

$$P(S = k) = \frac{n!}{k!\,(n-k)!} p^k \,(1-p)^{n-k} \quad \text{for } k = 0, \ldots, n.$$

This allows us to calculate the probability that S takes on any value k, as a function of the probability p. Since $P(S = k) = P(\widehat{p} = k/n)$, we can use the pmf to evaluate probabilities involving $\widehat{p}$ as well (from context, we are talking about the random estimator $\widehat{p}$). *In particular*, we can plug in the hypothesized value $p = 0.26$, pick a value c, and evaluate $P(|\widehat{p} - 0.26| \geq c)$, which is the probability that we deem a true hypothesis to be implausible. Deeming $p = 0.26$ to be implausible when it is true would be a mistake, and so we would like to ensure that this mistake is unlikely to happen.

The larger we make c, the less likely we are to deem a true hypothesis implausible. The standard way of picking c is to say that we would like to make this mistake with probability α or less, where α is a small number–like 0.005, 0.01, or 0.05–and then figure out what c must be to ensure this. Stated mathematically, we find the smallest c such that

$$P(|\widehat{p} - 0.26| \geq c) \leq \alpha.$$

As an example, suppose there are $n = 80$ students in the course. Then the event that $|\widehat{p} - 0.26| \geq c$ is equivalent to the event that $|S - 20.8| \geq 80c := d$. We'll do the calculation in terms of d and then convert back to c by dividing d by 80. Figure 2.1 shows the probabilities as a function of d. We can see that $d = 10.2$ is the smallest value for which the probability drops below 0.01. This means that we deem the hypothesis to be implausible if $|\widehat{p} - 0.26| \geq 10.2/80 = 0.1275$. In other words, if $\widehat{p}$ is between 0.1325 and 0.3875, we deem the hypothesis plausible, which might seem to be a surprisingly large range. This is because we picked a pretty stringent threshold probability of 0.01. A weaker threshold of 0.1 gives the range 0.1825 to 0.3375.

2.4 Null Hypothesis Significance Testing

The previous sections describe an example of null hypothesis significance testing (NHST). In this section, we give NHST a more formal treatment, including a more general view of decision rules and a discussion of statistical power. Once data $y_1, \ldots, y_n$ have been collected and a statistical model P_θ has been chosen, we define the null hypothesis as a statement about the unknown parameter θ, as in

$$H_0 : \theta = \theta_0.$$

It is possible to define more complicated null hypotheses, such as $\theta \geq \theta_0$, but we stick with the simple null hypothesis here. In our example above, the parameter was $\theta = p$, and the null hypothesis was $H_0 : p = 0.26$. The next step is to define a statistic t, which is a function of the data (hence the lowercase letter), and a decision rule, which says

$$\begin{array}{rl} \text{Fail to Reject } H_0 & \text{if } t \in A \\ \text{Reject } H_0 & \text{if } t \notin A. \end{array}$$

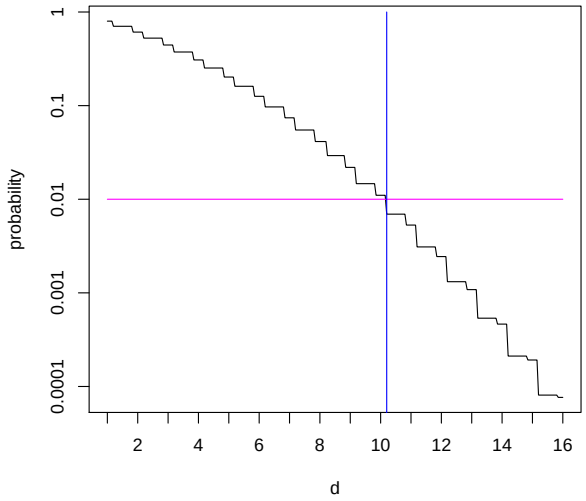


Figure 2.1: Plot of $P(|S - 20.8| > d)$ when $p = 0.26$ as a function of d. The smallest value of d for which the probability is less than 0.01 gives us our plausibility (significance) threshold. Here, that value is 10.2.

Rejecting H_0 is another way of saying that we deem H_0 to be implausible given the data. Failing to reject H_0 means that we deem it to be plausible given the data. In our example above, the statistic was $t = \widehat{p}$, and $A = (0.1325, 0.3375)$. Our decision rule is more general in that t need not represent an estimate of θ, and the set A need not be a single interval. For example, t could be a likelihood ratio or an F statistic. The choice of the set A depends on the significance level α of the test, and A is chosen so that

$$P_{\theta_0}(T \notin A) \leq \alpha.$$

Note that this probability is calculated assuming the null-hypothesized value θ_0, and we plug in the random variable version of the statistic T when we compute probabilities with the decision rule.

It should be noted that not all decision rules are created equally. While they all have the property that the probability of rejecting a true null hypothesis is less than α, some decision rules are better than others in the sense that their probability of rejecting a false null hypothesis could differ. Let θ_1 be a particular value of the parameter, and define

$$\text{power}(\theta_1) = P_{\theta_1}(T \notin A),$$

which is the probability that we reject the null hypothesis when it is false. Keep in mind that power calculations assume that the null is false but are calculated for decision rules determined under an assumption that the null is true. Read the previous sentence again and make sure you understand it.

Author's opinion: NHST has come under fire recently as one of many potential culprits in the reproducibility crisis, in which attempts to reproduce published scientific findings have often either failed or given results that are weaker than originally claimed. My intention here has been to give an explanation of the logic that undergirds NHST. My personal view on the matter is that the logical framework of NHST is sound but that any attempt to define hard "objective" thresholds on claims of discovery will invite people to game the system. This is why I think that, while statistical analyses are an important component of the scientific process, they should not be relied on as the sole determinant of whether data constitute a discovery. Statistical evidence is important but should be weighed against other forms of evidence.

2.5 Confidence Intervals

Hypothesis testing is useful in cases where the truthfulness of a particular claim is the primary goal of the analysis. Is the speed of light in a vacuum equal to 2.99792458×10^8 m/s? Does a drug have zero effect on patients' recovery time? However, in other cases, we simply want the analysis to return a range of plausible values of the parameter. Confidence intervals provide a logical framework for doing this.

Giving the textbook definition, a $(1-\alpha)$ confidence interval is a realization of a random interval that had probability $(1-\alpha)$ of containing the actual value of the parameter under the assumed statistical model for the data, regardless of the actual value of the parameter. The reason for defining the interval in terms of $(1-\alpha)$ will become clear soon.

Let's start with an absurd example to make sure we understand the definition. Suppose we have data $y_1, \ldots, y_{20}$, which we model as

$$Y_i \overset{iid}{\sim} N(\mu, \sigma^2),$$

a normal distribution with mean μ and variance σ^2. We construct a 0.95 confidence interval for μ as follows:

$$ci = \begin{cases} \emptyset & \text{if } y_1 \text{ is smallest} \\ (-\infty, \infty) & \text{otherwise,} \end{cases}$$

where $\emptyset$ is the empty set. No matter what the value of μ, it is not in the confidence interval if y_1 is smallest, and in the confidence interval if not. We use the statistical model to evaluate

$$P(\mu \in CI) = P(Y_1 \text{ is not smallest}) = 1 - P(Y_1 \text{ is smallest}) = 1 - 1/20 = 0.95.$$

This is true no matter what μ is, so ci is a valid confidence interval. In our dataset, suppose that y_8 is smallest. Then $ci = (-\infty, \infty)$, which gives absolutely no information about μ, but technically it is a 0.95 confidence interval.

We can use hypothesis tests to give more useful confidence intervals. Consider the set

$$ci = \{\mu^* | \text{we fail to reject } H_0 : \mu = \mu^* \text{ at level } \alpha\},$$

which should be read as the set of all parameter values μ^* such that we fail to reject $H_0 : \mu = \mu^*$. We could imagine doing a whole bunch of hypothesis tests at level α for different values of μ^*, and if we fail to reject, put that value into the confidence interval. In other words, the confidence interval (a set of plausible values) consists of all the parameter values that we can't rule out.

The derivation to show that ci is a $(1 - \alpha)$ confidence interval is confusingly simple. Let μ_0 represent the true value of the parameter. Remember we need to show that $P(\mu_0 \in CI) = 0.95$.

$$\begin{aligned} P(\mu_0 \in CI) &= P(\text{fail to reject } H_0 : \mu = \mu_0 \text{ at level } \alpha) \\ &= 1 - P(\text{reject } H_0 : \mu = \mu_0 \text{ at level } \alpha) \\ &= 1 - \alpha. \end{aligned}$$

The last equality is true because α is the precisely the probability that we reject a true null hypothesis! This method of constructing confidence intervals is sometimes called *inverting* the hypothesis test.

Suppose in the region of origin example that we get $s = 14$ when $n = 80$, and our decision rule is again to reject $H_0 : p = p_0$ when $|s - 80p_0| > d$. It takes a bit of computing, but we can do the hypothesis test for a grid of values for p_0 between 0 and 1 separated by 0.001. Collecting the values for which we fail to reject the null results in the confidence intervals in Figure 2.2. Note that intervals of higher confidence are longer, which happens because smaller α requires more evidence (larger d) to reject. Also, the confidence intervals are not symmetric around the estimate $\widehat{p} = 0.175$. This is because the variance of the sample proportion is $p(1-p)/n$, so values of p near $1/2$ have higher variance, so $p = 0.275$ is more likely to produce $\widehat{p} = 0.175$ than $p = 0.075$ is.

2.6 p-values

The p-value is closely related to the hypothesis test but has a slightly more stringent treatment of the decision rule. One must define an ordering of the statistics, that is, for any two values t_1 and t_2 of the statistic, we need to be able to say whether t_1 or t_2 is "more extreme." The decision rule in our region-of-origin example fits the mold because the decision rule was to reject $H_0 : p = 0.26$ if $t = |\widehat{p} - 0.26| \geq c$. The statistic is a positive number, and larger

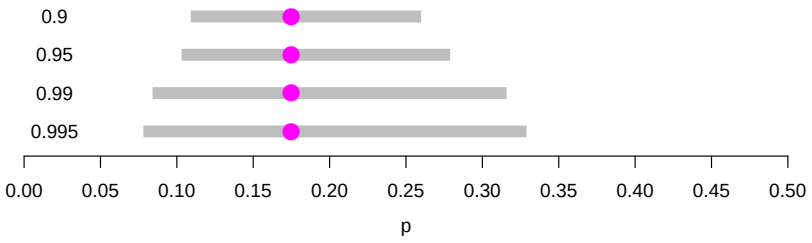


Figure 2.2: Confidence intervals for p in the region of origin example for four different confidence levels.

values correspond to larger deviations from the hypothesized proportion, and thus larger values are more extreme. The p-value is simply

$$\text{p-value} = P_{\theta_0}(T \geq t),$$

the probability–under the null hypothesis–of observing a statistic as extreme or more extreme than the one we did observe. This may look slightly confusing because both T, the random version of the statistic, and t, the actual nonrandom version, appear inside the probability. The probability calculation comes from the probability distribution of T.

Suppose we constructed a hypothesis test with significance α for a decision rule of the form above with $t_0 = c$. Conducting the test is equivalent to checking whether than p-value is greater or smaller than α. To see why this is true, suppose that a and b are two possible values of the statistic and that b is more extreme than a. Then the following inequality is true,

$$P(T \text{ more extreme than } b) \leq P(T \text{ more extreme than } a),$$

because if T is more extreme than b, it also has to be more extreme than a because b is more extreme than a. Suppose that the observed statistic t is less extreme than c. Then of course $P(T \geq t) \geq P(T \geq c) = \alpha$. On the other hand, if t is more extreme than c, $P(T \geq t) \leq P(T \geq c) = \alpha$. This means that if the p-value is less than α we reject, and if the p-value is greater than α we fail to reject.

Chapter 3

Linear Models and Regression

Joseph Guinness - BTRY 6020

Linear regression draws inferences about linear relationships between several variables and a response variable. It is probably the most used and most important statistical tool ever invented. In regression, we specify some variables as covariates (also known as predictors or features), and a single variable as the response variable. In this chapter, we cover simple linear regression, which has just a single covariate. Multiple covariates are covered in Chapter 3.

3.1 Least Squares and the t Distribution

Before discussing simple linear regression, let's review a few concepts from one-sample hypothesis testing. Suppose we have data $y_1, \ldots, y_n$ that we model as

$$Y_i \overset{iid}{\sim} N(b_0, \sigma^2).$$

Our goal is to make inferences about b_0 and σ^2. Focusing on b_0 and following the guidelines from the previous chapter, we must define a statistic or an estimate of b_0 and work out its sampling distribution in order to do the hypothesis test. To define an estimate, consider the residual sum of squares criterion (or least squares criterion)

$$rss(b_0^*) = \sum_{i=1}^{n} (y_i - b_0^*)^2,$$

which, for some arbitrary candidate parameter value b_0^*, is simply the sum of the squared deviations from the data to b_0^*. We would like our estimate of b_0 to

be close to the data y_i, so it makes sense that we would like $rss(\widehat{b}_0)$–the residual sum of squares of our estimate–to be small. Therefore, we simply define $\widehat{b}_0$ as

$$\widehat{b}_0 \text{ is the value of } b_0^* \text{ that minimizes } rss(b_0^*).$$

The least squares estimate can be a bit tough to swallow because it is defined as the minimizer of a criterion, or in other words, as the solution to a minimization problem. This is opposed to defining it with a formula. We have already seen this general idea once when we discussed decision rules for hypothesis tests. The cutoff c_α was defined as

$$c_\alpha \text{ is the smallest value } c \text{ such that } P(|\widehat{p} - p| \geq c) \leq \alpha.$$

Defining estimates as the solution to a problem is common, so it's best to get used to it now.

However, even though we define the estimate as the solution to a problem, in this case (and in some of the cases to follow), there is a simple formula for the parameter that minimizes rss–it's just the sample mean

$$\widehat{b}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

This makes it easy to know the sampling distribution of the random variable version of the estimate, also known as the estimator

$$\widehat{B}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N\Big(b_0, \frac{\sigma^2}{n}\Big).$$

This doesn't tell us everything we need to know about the sampling distribution, because we don't know what σ^2 is. If we did, then could define a decision rule such as

$$\text{reject } H_0 : b_0 = b_0^* \quad \text{if} \quad |\widehat{b} - b_0^*| \geq c,$$

and work out what c needs to be in order that the false positive probability is equal to our chosen significance level α.

Since we don't know σ^2, we need to get some estimate of σ^2. We use

$$\widehat{\sigma}^2 = \frac{rss(\widehat{b}_0)}{n-1}.$$

We'll discuss later in this chapter why we choose this estimate, so just take it for granted now. Then we can use the sampling distribution $\widehat{B}_0 \sim N(b_0, \widehat{\sigma}^2/n)$ to find the cutoff c and conduct our hypothesis tests. This is how things were done before 1908. However, this isn't quite right because the sampling distribution includes a parameter $\widehat{\sigma}^2$ that depends on the data, which is a no-no because repeated experiments would all have different sampling distributions. Instead,

in a 1908 paper, William Sealy Gosset proposed the alternative statistic (and its RV version)

$$t = \frac{\widehat{b}_0 - b_0^*}{\widehat{\sigma}/\sqrt{n}}, \quad T = \frac{\widehat{B}_0 - b_0^*}{\widehat{\sigma}/\sqrt{n}}.$$

Caution: this is one of those situations I've warned you about, where the symbol $\widehat{\sigma}$ refers to both a data-dependent quantity (in the definition of t) and a RV (in the definition of T). Gosset noticed that the distribution of T does not depend on the hypothesized parameter value b_0^*, nor does it depend on the other unknown variance parameter σ^2. It only depends on n, and we refer to T_{n-1} as a random variable with a t distribution with $n-1$ degrees of freedom (more on degrees of freedom later). Its distribution was also equal to one that Karl Pearson had derived earlier for other purposes.

These mathematical results were a big advance because they allow us to do hypothesis tests based directly on the statistic t and its sampling distribution, without using any approximations related to plugging in estimated values of the parmaeters. Pearson's distribution was renamed as the Student t distribution, in honor of the pseudonym "Student" that Gosset had to use when publishing his mathematical work. Gosset used a pseudonym in order to give some protection from revealing trade secrets of his employer, Guinness brewing, where Gosset later went on to become Master Brewer (equivalent to CEO) for a short period.

3.2 Simple Linear Models

Figure 3.1 contains a subset of Francis Galton's height data. Galton collected the data because he was interested in studying the nature of genetic heritability. The figure contains a sample data table and a scatterplot of the dataset. The data table has three columns. The first column contains the labels assigned to each family. The second column is midparent height (average of parents' height adjusted for sex), and the third is the height of one female child. Each row contains data from one family. The scatterplot has midparent height on the horizontal axis and child height on the vertical axis. This particular dataset has $n = 205$ families. In mathematical notation, we write x_i for midparent height in family i, y_i for child height from family i, and i runs from 1 up to 205.

A linear model for y_i in terms of x_i means that we model y_i with a random variable Y_i whose expected value $E(Y_i)$ is a linear function of x_i. Namely,

$$E(Y_i) = b_0 + b_1 x_i. \tag{3.1}$$

Note that b_0 now has a different interpretation than it did in the previous section, where it was the expected value of Y_i. Here, b_0 is the expected value of Y_i when $x_i = 0$. This is an important distinction. In linear models, the interpretations

i	x_i	y_i
1	75.43	69
2	73.66	65.5
3	75.43	69
4	72.06	63
5	69.09	62.5
6	69.09	62.5
7	73.72	64
8	72.91	68
9	73.66	65.5
10	73.66	73.5
11	70.48	66
12	73.72	64
13	69.09	62.5
14	72.91	70.5
15	71.6	64.5
⋮	⋮	⋮

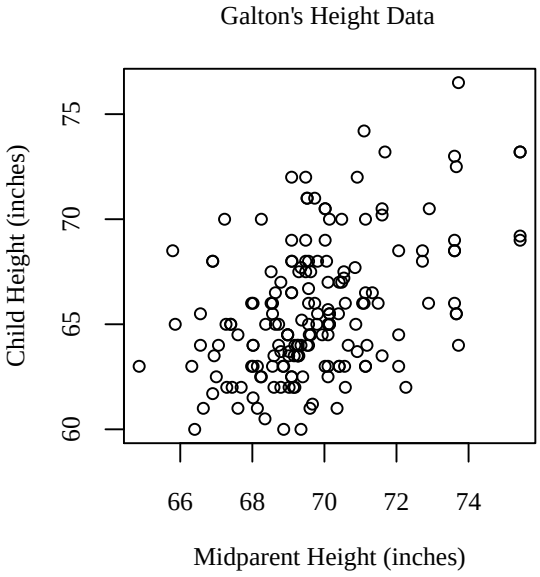


Figure 3.1: Galton's height data

of individual parameters often change when we add or subtract terms from the model, so keep this in mind when interpreting your results.

Of course, Equation (3.1) does not constitute a statistical model because it does not specify the entire probability distribution for Y_i, only its expected value. The normal linear model is

$$Y_i = b_0 + b_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{3.2}$$

which says that Y_i is equal to its expectation plus an independent normal random variable with variance σ^2. We will refer to x_i as the covariate, y_i and Y_i as the response, and ε_i as the error term. We are calling this the normal linear model, but to be complete, we should really call it the normal, independent, equal variance linear model. We'll never finish the course if we force ourselves to say all that, so let's drop it and move on with our lives. This is not to say that these assumptions are unimportant; assuming independence when it's not true is one of the worst data analysis mistakes you can make. Later on in the book we will consider normal linear models that are not independent or equal variance.

Galton's research was motivated by that of his more famous half cousin, Charles Darwin. In particular he wanted to study population stability in the presence of heritable traits. A mathematical analysis of the following model

$$Y_i = x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

reveals that it leads to unstable populations, in the sense that the population variance of height would increase over multiple generations. Galton wanted to consider more flexible models and proposed the linear model as an alternative that leads to stable populations. The unstable model is a special case of our linear model corresponding to $b_0 = 0$ and $b_1 = 1$.

Our objective is to use Galton's data to make inferences about the unknown parameters b_0, b_1, and σ^2. To achieve this, we will define estimates of these parameters and use the statistical model to work out their sampling distributions. To estimate b_0 and b_1, we use the residual sum of squares criterion (a.k.a. the least squares criterion),

$$rss(b_0^*, b_1^*) = \sum_{i=1}^{n} (y_i - b_0^* - b_1^* x_i)^2$$

and select estimates $\widehat{b}_0$ and $\widehat{b}_1$ as the values of b_0^* and b_1^* that minimize the residual sum of squares criterion. This means that

$$rss(\widehat{b}_0, \widehat{b}_1) \leq rss(b_0^*, b_1^*) \quad \text{for any other values } b_0^* \text{ and } b_1^*.$$

Let's take a moment to review the meaning that we've given to our notation.

Notation	meaning
b_0 and b_1	"true" unknown values of the parameters
$\widehat{b}_0$ and $\widehat{b}_1$	estimates of the parameters from data
b_0^* and b_1^*	no meaning, arbitrary alternative values of the parameters

However, in this case, even though the parameter estimates are defined as the solution to the least squares problem, they also have easy formulas. The formulas are

$$\widehat{b}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} := \frac{sxy}{sxx} \qquad \widehat{b}_0 = \overline{y} - \widehat{b}_1 \overline{x},$$

where $\overline{y}$ is the sample mean of $y_1, \ldots, y_n$, and $\overline{x}$ is the sample mean of $x_1, \ldots, x_n$.

The estimators are the random variable version of the estimates,

$$\widehat{B}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad \widehat{B}_0 = \overline{Y} - \widehat{B}_1 \overline{x}.$$

The probability distributions of the estimators is known as the sampling distribution. The estimator $\widehat{B}_1$ can be expressed as a linear combination of $Y_1, \ldots, Y_n$,

$$\widehat{B}_1 = \sum_{i=1}^{n} v_i Y_i, \quad \text{where} \quad v_i = \frac{(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

A linear combination of $Y_1, \ldots, Y_n$ means that you can simply write the formula as the sum of each Y_i multiplied by a nonrandom coefficient v_i. This implies that $\widehat{B}_1$ and $\widehat{B}_2$ are both normal, and further calculations show that

$$\widehat{B}_0 \sim N\Big(b_0,\, \sigma^2\Big(\frac{1}{n} + \frac{(\overline{x})^2}{sxx}\Big)\Big) \quad \text{and} \quad \widehat{B}_1 \sim N\Big(b_1,\, \frac{\sigma^2}{sxx}\Big)$$

Both $\widehat{B}_0$ and $\widehat{B}_1$ are **unbiased**, which means that their expected values are equal to the quantities that they aim to estimate (b_0 and b_1). The covariance between the estimators is

$$\mathrm{Cov}(\widehat{B}_0, \widehat{B}_1) = -\sigma^2 \frac{\overline{x}}{sxx},$$

which means that if $\overline{x} \neq 0$, the estimators are dependent.

The other parameter in the model is $\sigma^2 = \mathrm{Var}(\varepsilon_i)$. If we knew the true linear parameter values, we could recover ε_i as

$$\varepsilon_i = Y_i - b_0 - b_1 x_i,$$

and then estimate σ^2 by computing

$$\frac{rss(b_0, b_1)}{n} = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - b_0 - b_1 x_i\big)^2.$$

This estimate would be unbiased, but we can't use it because we don't know b_0 and b_1. We do have estimates $\widehat{b}_0$ and $\widehat{b}_1$, so we can substitute the estimated values and use the estimate

$$\frac{rss(\widehat{b}_0, \widehat{b}_1)}{n}.$$

However, we know for sure that $rss(\widehat{b}_0, \widehat{b}_1) \leq rss(b_0, b_1)$, and so if our estimate based on $rss(b_0, b_1)$ is unbiased, then an estimate based on $rss(\widehat{b}_0, \widehat{b}_1)$ must be biased downwards (too small). The solution is to adjust the estimate and use

$$\widehat{\sigma}^2 = \frac{rss(\widehat{b}_0, \widehat{b}_1)}{n-2}$$

instead. This estimate of σ^2 is unbiased.

3.3 Degrees of Freedom

Why $n-2$ in the denominator? We can verify mathematically that $n-2$ is the right quantity to put in the denominator, by showing that $\widehat{\sigma}^2$ is unbiased. But we don't want to have to do that for every different model we try. The general

solution for what to put in the denominator is determined by the **residual degrees of freedom**. Degrees of freedom (dof) is a somewhat tricky concept, so let's try to demystify it.

Most of the models we will consider this semester can be written in the form

$$Y_i \sim N(\mu_i, \sigma^2),$$

where μ_i describes the **mean** part of the model, and σ^2 describes the residual **variance** part of the model.

Definition 1 *The **model degrees of freedom** is the number of numbers required to specify the mean part of the model.*

In simple linear regression, we have $\mu_i = b_0 + b_1 x_i$, which describes an unconstrained line. Unconstrained lines can be specified several ways. Here, we use the intercept and slope, but if you think back to our first algebra course, we could have specified it in point-slope form, which means that you give me an x value, and I have to tell you the corresponding y value and the slope, or in point-point form, which means that you give me two x values, and I have to give you the two corresponding y values. In each specification, you need two pieces of information from me in order to know what the line is. Thus, the model degrees of freedom for the simple linear model is two. This happens to be equal to the number of parameters in the mean part of the model, but we will see later that the model degrees of freedom is not always equal to the number of parameters in the mean.

Definition 2 ***residual dof*** $= n -$ *model dof*

The residual degrees of freedom is simply n minus the model degrees of freedom. It has an interpretation as the number of degrees of freedom left over in the data after accounting for the degrees of freedom in the model (whose parameters were estimated by minimizing rss). We use the degrees of freedom from the residual to estimate the variance parameter σ^2, and so

$$\widehat{\sigma}^2 = \frac{rss(\widehat{\boldsymbol{b}})}{\text{residual dof}},$$

where $\widehat{\boldsymbol{b}} = (\widehat{b}_0, \widehat{b}_1, \ldots, \widehat{b}_p)$.

3.4 t Distributions for Simple Linear Model

We use the term **standard error** to refer to the square root of the variance of an estimator (a.k.a. the estimator's standard deviation). To me, it's a bit confusing to have multiple words meaning the same thing, but "standard error" is so commonly used that we will use it here, too. For the simple linear model,

the standard errors are

$$\text{se}(\widehat{B}_0) = \sqrt{\text{var}(\widehat{B}_0)} = \sqrt{\sigma^2\Big(\frac{1}{n} + \frac{(\overline{x})^2}{sxx}\Big)}$$
$$\text{se}(\widehat{B}_1) = \sqrt{\text{var}(\widehat{B}_1)} = \sqrt{\frac{\sigma^2}{sxx}}.$$

These formulas for the standard errors include the unknown variance parameter σ^2. We use the notation $\widehat{\text{se}}(\widehat{B}_j)$ to refer to the standard errors with $\widehat{\sigma}$ plugged in for σ, and $\widehat{\text{SE}}(\widehat{B}_j)$ to refer to standard errors with the RV version of $\widehat{\sigma}$ plugged in.

For the simple linear model, we can test the hypotheses

$$H_0 : b_0 = b_0^* \quad \text{or} \quad H_0 : b_1 = b_1^*,$$

by defining decision rules of the form

$$\text{reject } H_0 \quad \text{if} \quad |t| := \left|\frac{\widehat{b}_j - b_j^*}{\widehat{\text{se}}(\widehat{B}_j)}\right| > c,$$

and using the sampling distribution of T. Fortunately, its sampling distribution is known to be a t distribution with $n-2$ degrees of freedom, so we can easily do calculations in R.

Chapter 4

Multiple Linear Models

Joseph Guinness - BTRY 6020

We often want to determine the relationship between a response variable and several covariates. We could estimate separate simple linear models for each of the individual covariates, but this would miss potentially very useful information. For example, in the lecture, we will explore a dataset that contains biometric measurements of developing children. The results will show that both covariates, height at age 9 and leg length at age 9, are individually positively correlated with the response, height at age 18. However, the story is more complicated when we analyze the simultaneous relationship between the two covariates and the response.

The multiple linear model helps us answer the question, "what is the effect of one variable on the response after controlling for other variables." What is the effect of leg length at age 9 after controlling for height? A simultaneous analysis of several variables helps us understand how one covariate affects the response given a fixed value of the other covariate. Multiple linear regression is a technique for simultaneously estimating the effects.

As before, let $i = 1, \ldots, n$ be labels for the observations, and $y_1, \ldots, y_n$ be the response variables. Since we have multiple covariates, we need to make a small change to our notation for the covariates. Let x_{ij} be the value of the jth covariate for the ith subject. For example, if we had two covariates, our data table might look something like this:

i	y_i	x_{i1}	x_{i2}
1	y_1	x_{11}	x_{12}
2	y_2	x_{21}	x_{22}
3	y_3	x_{31}	x_{32}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}

The multiple linear model can be written in several different equivalent ways,

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots b_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$
$$Y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + \varepsilon_i$$
$$Y_i = \sum_{j=0}^{p} b_j x_{ij} + \varepsilon_i,$$

where $b_0, \ldots, b_p$ and σ^2 are unknown parameters. In the last expression, we define $x_{i0} = 1$. No matter how we write the model, $E(Y_i)$ is a linear function of each of the covariates, in the sense that if we increase x_{ij} by 1 while holding all other covariates constant, $E(Y_i)$ increases by b_j. The intercept parameter b_0 is interpreted as $E(Y_i)$ when all covariates are zero. In our notation, there are p covariates (I don't know who decided to use p). Adding the intercept, the multiple linear model has $p + 1$ model degrees of freedom.

Parameter estimation and testing in the multiple linear model is basically the same as it was in the simple linear model, and will be the same for most of the models we encounter this semester. We use the residual sum of squares criterion to estimate the model parameters. Define $\boldsymbol{b} = (b_0, \ldots, b_p)$ to be the vector of parameters, and $\boldsymbol{b}^*$ to be an arbitrary choice for the parameters. The concept of a vector is nothing to be afraid of; for our purposes, it's just an ordered list of numbers. The least squares estimate

$$\widehat{\boldsymbol{b}} \text{ is the value of } \boldsymbol{b}^* \text{ that minimizes } rss(\boldsymbol{b}^*) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} b_j^* x_{ij} \right)^2.$$

In simple linear regression, there was a relatively simple formula for $\widehat{\boldsymbol{b}}$ in terms of sxy and sxx. In multiple linear regression, there is a corresponding formula involving matrix multiplication and matrix inverses. In the interest of time and space, we'll skip the formula and simply view $\widehat{\boldsymbol{b}}$ as the minimizer of our rss criterion.

The sampling distribution of $\widehat{B}_j$, the random variable version of $\widehat{b}_j$, is normal and unbiased for b_j. No surprises there. Both the variance of $\widehat{B}_j$ and $\text{cov}(\widehat{B}_j, \widehat{B}_k)$ have known but complicated formulas that we'll also skip in the interest of time.

We estimate the error variance with

$$\widehat{\sigma}^2 = \frac{rss(\widehat{\boldsymbol{b}})}{n-(p+1)}$$

because there are $n-(p+1)$ residual degrees of freeom. Be careful when you write the dof; $n-(p+1)$ is not the same as $n-p+1$ but is the same as $n-p-1$. We use the same notation $se(\widehat{B}_j)$, $\widehat{se}(\widehat{B}_j)$, and $\widehat{SE}(\widehat{B}_j)$ to indicate whether the sampling distributions use the true or estimated error variance.

Testing involves the t distribution. In order to test an individual null hypothesis like $H_0 : b_j = 3.2$ (or any other number), we use the test statistic

$$t = \frac{\widehat{b}_j - 3.2}{\widehat{se}(\widehat{B}_j)},$$

whose random variable version has has a t distribution with $n-p-1$ degrees of freedom. Sometimes we are interested in hypothesis tests that involve individual parameters, and at other terms we wish to test a hypothesis like $H_0 : (b_1 = 0 \text{ and } b_2 = 0)$. Don't ask the t-test such complicated questions. We'll need to consult our friend the F-test for that. The F-test will be covered in a later chapter.

Most of the models we study this semester can be written as multiple linear models, and therefore we can use the residual sum of squares to estimate the parameters, and t tests to conduct inference. The trick is figuring out how to take our conceptual idea for our model and figure out how to place it within the framework of a multiple linear model.

As an example, suppose we had a single covariate x_i, but instead of the simple linear model, in which the expected response is a linear function of x_i, we wanted to fit a quadratic model:

$$Y_i = b_0 + b_1 x_i + b_2 x_i^2 + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The conceptual idea is that we have a single covariate x_i, and we believe that the expected response is a quadratic function of the covariate. Since the expected response in this model is not linear in x_i, it does not fit into our simple linear model framework. It doesn't immediately look like it fits into our multiple linear model framework either, since there is only one covariate rather than multiple. However, we can use the clever trick of defining our own covariates. Let $x_{i1} = x_i$, and let $x_{i2} = x_i^2$. This means that our data table looks like this:

which does fit into the framework. This allows us to do t tests on b_0, b_1, and b_2. In particular, we can do a t test of $H_0 : b_2 = 0$, which compares the quadratic model, which has $b_2 \neq 0$, to the simple linear model, which has $b_2 = 0$. This allows us to test whether we have enough evidence to reject a simple linear model in favor of a quadratic model.

i	y_i	x_{i1}	x_{i2}
1	y_1	x_1	x_1^2
2	y_2	x_2	x_2^2
3	y_3	x_3	x_3^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n	x_n^2

4.1 Berkeley Guidance Study Data

The Berkeley Guidance Study data comes from children born in 1928 and 1928 in Berkeley, CA. We will look at the girls data, which can be read into R using the command `data("BGSgirls",package="alr4")`. Here is a sample of the data:

HT2	HT9	LG9	HT18
87.70	133.40	28.40	158.90
90.00	134.80	26.90	166.00
89.60	141.50	31.90	162.20
90.30	137.10	31.80	167.80
89.40	136.10	27.70	170.90
85.50	130.60	23.40	164.90
90.20	136.00	27.20	168.10
82.20	128.00	25.10	164.00
85.60	132.40	27.50	163.30
97.30	152.50	32.70	183.20

Each row corresponds to data from one girl. The "HT" variables are heights, and "LG" is leg circumference. The numbers following the variable indicate the age at which the measurement was taken. We are interested in understanding the relationship between height at age 18 (HT18) and the other three variables.

First, we define our mathematical notation for the data.

$$\begin{aligned} y_i &= \text{height at age 18 of girl } i \\ x_{i1} &= \text{height at age 2 of girl } i \\ x_{i2} &= \text{height at age 9 of girl } i \\ x_{i3} &= \text{leg circ. at age 9 of girl } i \end{aligned}$$

Figure 4.1 contains scatterplots showing the relatinship between each of the covariates and the response, and we can see that they are all positively correlated with the response. Unsurprisingly, height at age 9 has the strongest relationship. Let's fit two simple linear models, one using height at age 2, and the other using

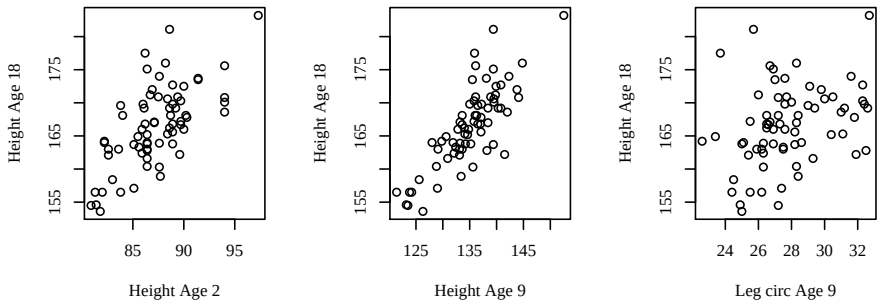


Figure 4.1:

height at age 9 as the covariate, to confirm what we see visually in the plot:

$$\text{Model 1:} \quad Y_i = b_0 + b_1 x_{i1} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$
$$\text{Model 2:} \quad Y_i = b_0 + b_1 x_{i2} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Note that we are re-using some notation, so the symbols have different meanings depending on which model they appear in. In Model 1, b_1 is the expected increase in height at age 18 when height at age 2 increases by 1 cm, whereas in Model 2, b_1 is the expected increase in height at age 18 when height at age 9 increases by 1 cm.

Here are abbreviated summaries of the model fits:

```
> summary(m1)
Call:
lm(formula = HT18 ~ HT2, data = BGSgirls)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.9748    14.4528   4.219 7.44e-05 ***
HT2           1.2099     0.1655   7.310 3.92e-10 ***

Residual standard error: 4.579 on 68 degrees of freedom
Multiple R-squared:   0.44,Adjusted R-squared:  0.4318

> summary(m2)
Call:
lm(formula = HT18 ~ HT9, data = BGSgirls)

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.40628   10.46260   4.627 1.72e-05 ***
HT9          0.87432    0.07737  11.301  < 2e-16 ***

Residual standard error: 3.607 on 68 degrees of freedom
Multiple R-squared:  0.6526,Adjusted R-squared:  0.6474
```

We see that in both models, we have strong evidence that b_1 is not equal to zero, judging by the large t statistics and the small p-values. We estimate that increasing age 2 height by 1 cm increases the expected age 18 height by 1.2 cm, and increasing the age 9 height by 1 cm increases the age 18 height by 0.87 cm. Should we conclude that age 2 height is a better predictor of age 18 height because it has a larger coefficient?

The answer is no, so think about why b_1 should be larger in the age 2 model. Judging by the larger R-squared value (0.6526 vs. 0.44) and smaller estimate of σ (3.607 vs. 4.579), age 9 height is a better predictor of age 18 height. This confirms what we saw in the scatterplots.

Next we are going to put these two variables together in the same model, but before we do that, let's think about growth and heights by considering a simple example. Suppose that the following is true:

$$\begin{aligned} x_{11} &= 85 \quad x_{12} = 135 \\ x_{21} &= 86 \quad x_{22} = 135 \\ x_{31} &= 85 \quad x_{32} = 135 \\ x_{41} &= 85 \quad x_{42} = 136 \end{aligned}$$

We have four girls, the first two differ in age 2 height but not age 9 height, and the second two differ in the age 9 height but not age 2 height? How do we expect that their age 18 heights will differ? Which of the four girls would you bet was shortest tallest at age 18? Which would be tallest? Take a moment to ponder these questions before reading on.

Now, let's put both variables together in the same model:

$$\text{Model 3:} \quad Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

The coefficients have a different meaning in this model. Sticking with our example above, we can isolate coefficients by looking at differences in expected values of the responses:

$$\begin{aligned} b_1 &= E(Y_2) - E(Y_1) \\ b_2 &= E(Y_4) - E(Y_3). \end{aligned}$$

If these equations aren't obvious to you, write out the expected values of the Y_i variables, plugging in the values of x_{i1} and x_{i2} from the example above.

b_1 is the expected difference in age 18 heights for two girls that differ in the age 2 height by 1 cm but have the same age 9 height. And b_2 is the expected difference in age 18 height for two girls that differ in age 9 height by 1 cm but have the same age 2 height. In other words, b_1 is the effect of age 2 height, holding age 9 height constant, and b_2 is the effect of age 9 height, holding age 2 height constant.

Here is the R output from fitting Model 3:

```
Call:
lm(formula = HT18 ~ HT2 + HT9, data = BGSgirls)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8022 -1.5632 -0.2918  1.9155 10.9553

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.8817    11.7030   3.493  0.00085 ***
HT2           0.2682     0.1920   1.397  0.16699
HT9           0.7568     0.1139   6.643 6.56e-09 ***

Residual standard error: 3.582 on 67 degrees of freedom
Multiple R-squared:  0.6624,Adjusted R-squared:  0.6523
```

We have very strong evidence that age 18 height depends on age 9 height after controlling for age 2 height, but do not have much evidence that age 18 height depends on age 2 height after controlling for age 9 height. Moreover, the estimated coefficient age 2 height ($\widehat{b}_1 = 0.2682$) is much smaller than the estimate coefficient for age 9 height ($\widehat{b}_2 = 0.7568$), even though there is more variation in age 9 height than age 2 height. Do these results confirm you thinking based on the example with the four girls above? If not, where did your reasoning go wrong?

We intepret this to mean that age 2 height provides little additional information about age 18 height after knowing age 9 height. This makes a lot of sense; if I told you a child's age 9 height, and you wanted to predict their age 18 height, then you wouldn't ask what their age 2 height was. That information is irrelevant after knowing the age 9 height.

What about leg circumference? Consider another model:

$$\text{Model 4:} \quad Y_i = b_0 + b_1 x_{i2} + b_2 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

This model covariates age 9 height and age 9 leg circumference, and is designed to answer the question, "does age 18 height depend on age 9 leg circumference after controlling for age 9 height?"

Here are the results of that fit:

```
> summary(m4)
Call:
lm(formula = HT18 ~ HT9 + LG9, data = BGSgirls)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.07628   10.11270   4.359 4.61e-05 ***
HT9          1.02619    0.09217  11.133  < 2e-16 ***
LG9         -0.58165    0.21114  -2.755  0.00755 **

Residual standard error: 3.444 on 67 degrees of freedom
Multiple R-squared:  0.6879,Adjusted R-squared:  0.6786
```

Are the results surprising to you? Could we have made an error? If not, how do you intepret them? Can you think of a good reason for why the LG9 coefficient should be negative?

Chapter 5

The F Test

Joseph Guinness - BTRY 6020

We have now introduced the multiple linear model, provided interpretations for its parameters, and learned about the t-test as a tool for making inferences about individual mean parameters. We have also hinted that the multiple linear model is capable of modeling a wide range of types of datasets. We will spend a substantial amount of time exploring how to adapt the multiple linear model to handle categorical covariates and interaction effects. The questions we hope to answer by analyzing data will be more complex and involve hypothesis tests that involve multiple parameters simulataneously.

For example, if we have three covariates and model the responses y_i as

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

we may be interested in a hypothesis of the form

$$H_0 : (b_2 = 0 \text{ and } b_3 = 0).$$

This might come up if y_i is height at age 18, x_{i1} is height at age 9, x_{i2} is height at age 2, and x_{i3} is leg circumference at age 9. The hypotheses asks whether *either* height at age 2 *or* leg circumference at age 9 contain any linear information about height at age 18 beyond that which is contained in height at age 9.

Or we might be interested in a hypotheses of the form

$$H_0 : b_1 = b_2 = b_3.$$

This might be of interest if we were modeling the heights of the youngest child in a family of 4, and x_{i1} is the height of the oldest child, x_{i2} the second-oldest, and x_{i3} the third-oldest (adjusted for sex). This hypothesis asks whether heights of siblings all contribute equally, or if children born closer or further away in order provide different amounts of information about the youngest child's height.

The t-test is not capable of testing these types of hypotheses that involve multiple parameters (sidenote: you could devise a t-test for the hypothesis $b_1 = b_2$ but not $b_1 = b_2 = b_3$). However, the F-test can.

Put simply, the F-test is a comparison of two competing models for the data. In the F-test, one of the models must be a special case of the other model, in other words, a simplified version of the other model. This is usually achieved by specifying a *full model* and a *reduced model*, which is obtained by placing some kind of restriction on the parameters from the full model. This is most easily explained by looking at some examples.

Consider again our heights example. If we wanted to test the hypothesis that neither age 2 height nor age 9 leg size contained additional information about age 18 height, we would write down a full model with all three variables:

$$\text{Full Model:} \quad Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

where y_i is height at age 18, x_{i1} is height at age 9, x_{i2} is height at age 2, and x_{i3} is leg circumference at age 9. Then we would compare the full model to the reduced model

$$\text{Reduced Model:} \quad Y_i = b_0 + b_1 x_{i1} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

which has only age 9 height as a predictor. This qualifies as a reduced model, since it corresponds to setting b_2 and b_3 to zero in the full model. Thus it is a special case of the full model. Comparing these two models is equivalent to testing the hypothesis

$$H_0 : (b_2 = 0 \text{ and } b_3 = 0),$$

which we said that the t-test is not capable of testing.

The procedure for conducting an F-test is pretty straightforward.

1. Fit the full model and record $rss1$ and $df1$, the residual sum of squares and residual degrees of freedom from the full model.
2. Fit the reduced model and record $rss0$ and $df0$, the residual sum of squares and the residual degrees of freedom from the reduced model.
3. Compute
$$f = \frac{(rss0 - rss1)/(df0 - df1)}{rss1/df1}.$$
4. Compare f to an F distribution with $df0 - df1$ numerator degrees of freedom and $df1$ denominator degrees of freedom.

There are R functions to do all of these calculations for you. However, let's unpack the formula for the f statistic to give some intuition for what it means.

First, due to the fact that the reduced model is a special case of the full model, $rss1$ must be less than or equal to $rss0$. To see why, consider the simple linear model and its special case, the model with no covariates:

$$\begin{aligned} \text{Full Model:} \quad & Y_i = b_0 + b_1 x_i + \varepsilon_i \\ \text{Reduced Model:} \quad & Y_i = b_0 + \varepsilon_i. \end{aligned}$$

The least squares estimates of b_0 and b_1 ($\widehat{b}_0$ and $\widehat{b}_1$) are the values of the parameters that minimize the residual sum of squares criterion in the full model. In particular,

$$rss1 = rss(\widehat{b}_0, \widehat{b}_1) \leq rss(\overline{y}, 0) = rss0.$$

The relationship $rss(\overline{y}, 0) = rss0$ is true because $\overline{y}$ is the least squares estimate of b_0 in the reduced model. The $\leq$ relationship is true because $\widehat{b}_0$ and $\widehat{b}_1$ minimize the residual sum of squares over all potential candidates for b_0 and b_1, and $\overline{y}$ and 0 are two valid potential candidates for b_0 and b_1.

Second, the full model has a larger model degrees of freedom than the reduced model–by definition the reduced model is a less flexible version of the full model. Since residual degrees of freedom is n minus model degrees of freedom, $df0$ must be larger than $df1$.

Taken together with our inequality $rss1 \leq rss0$, the f statistic must be a positive number. The numerator is a measure of how much the residual sum of squares is reduced by considering the more flexible full model. The reduction of rss is scaled by the additional degrees of freedom in the full model ($df0 - df1$). Scaling by the additional model complexity makes some sense because we generally want to balance model fit (rss) against model complexity (model dof).

The denominator is the estimate of the error variance taken from the full model. So the scaled reduction in rss is itself scaled by the error variance. This also makes some sense. The reduction in rss depends on the units of measurement; if we measured height in inches rather than centimeters, we would get a different reduction in rss. However, if we scale by the error variance, the f statistic becomes a unitless quantity, and thus not dependent on the units of measurement.

The F-test is also capable of conducting hypothesis tests such as

$$H_0 : b_1 = b_2 = b_3$$

from the model

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \varepsilon_i.$$

We simply set the full model to the above model and the reduced model to

$$\begin{aligned} Y_i &= b_0 + b_1 x_{i1} + b_1 x_{i2} + b_1 x_{i3} + \varepsilon_i \\ Y_i &= b_0 + b_1 (x_{i1} + x_{i2} + x_{i3}) + \varepsilon_i, \end{aligned}$$

which is a special case of the full model that codifies our null hypothesis that b_1, b_2, and b_3 are all equal.

5.1 Berkeley Girls Data

The F-test is most commonly used for categorical (factor) models, which we will study in the coming chapters. But it applies equally well to models with continuous (numeric) covariates. Let's consider some additional covariates from the Berkeley Guidance study data:

$$\begin{aligned} y_i &= \text{ height at age 18} \\ x_{i1} &= \text{ height at age 2} \\ x_{i2} &= \text{ weight at age 2} \\ x_{i3} &= \text{ height at age 9} \\ x_{i4} &= \text{ weight at age 9} \\ x_{i5} &= \text{ leg circumference at age 9} \\ x_{i6} &= \text{ strength at age 9} \end{aligned}$$

The following model contains all of these covariates

$$Y_i = b_0 + \sum_{j=1}^{6} b_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

Here is the R output from fitting that model:

```
Call:
lm(formula = HT18 ~ WT2 + HT2 + WT9 + HT9 + LG9 + ST9, data = BGSgirls)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.68358   15.92247   0.294   0.7696
WT2          0.34648    0.42840   0.809   0.4217
HT2          0.10665    0.20396   0.523   0.6029
WT9         -0.48533    0.20656  -2.350   0.0219 *
HT9          1.19857    0.15438   7.763 9.25e-11 ***
LG9          0.20721    0.39517   0.524   0.6019
ST9         -0.07041    0.03443  -2.045   0.0451 *

Residual standard error: 3.274 on 63 degrees of freedom
Multiple R-squared:  0.7348,Adjusted R-squared:  0.7095
F-statistic: 29.09 on 6 and 63 DF,  p-value: < 2.2e-16
```

Let's first focus on the age 2 variables. The t-test for WT2 tests whether the response depends linearly on weight at age 2 after controlling for all other variables in the model. The test has a p-value of 0.4217, which does not provide any evidence against the hypothesis that they are unrelated given the other variables. We see a similar story for height at age 2.

Suppose we wanted to test whether the response is linearly related to either of the age 2 variables after controlling for the age 9 variables. This is a different

test than the t-tests above because we are not controlling for the other age 2 variable. The hypothesis is

$$H_0 : b_1 = 0 \text{ and } b_2 = 0$$

We need an F-test for this hypothesis. The reduced and full models are

$$\text{Reduced: } Y_i = b_0 + \sum_{j=3}^{6} b_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

$$\text{Full: } Y_i = b_0 + \sum_{j=1}^{6} b_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

Note the difference in the range of the sum over j. We have already fit the full model. When we fit the reduced model, we get

```
Call:
lm(formula = HT18 ~ WT9 + HT9 + LG9 + ST9, data = BGSgirls)

Residuals:
   Min     1Q Median     3Q    Max
-7.543 -1.962  0.085  1.806  8.299

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.05371   15.69040   0.450   0.6545
WT9         -0.45569    0.18959  -2.403   0.0191 *
HT9          1.26779    0.11746  10.794 3.91e-16 ***
LG9          0.24458    0.38691   0.632   0.5295
ST9         -0.06967    0.03283  -2.122   0.0376 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.266 on 65 degrees of freedom
Multiple R-squared:  0.7276,Adjusted R-squared:  0.7109
F-statistic: 43.42 on 4 and 65 DF,  p-value: < 2.2e-16
```

Assuming we have saved the full model as `m1` and the reduced model as `m2`, we can calculate the F statistic and its p-value as follows:

```
> rss0 <- sum(m2$residuals^2)
> df0 <- m2$df.residual
> rss1 <- sum(m1$residuals^2)
> df1 <- m1$df.residual
> fstat <- (rss0-rss1)/(df0-df1)/(rss1/df1)
> fstat
[1] 0.8467008
> 1-pf(fstat,df0-df1,df1)
```

```
[1] 0.4336484
```

Not surprisingly, we fail to reject our hypothesis.

Let's dig a little deeper. Suppose we wanted to test the hypothesis that there is no linear relationship between the response and the non-height age 9 variables after controlling for age 9 height. Then the reduced and full models are

$$\text{Reduced: } Y_i = b_0 + b_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

$$\text{Full: } Y_i = b_0 + \sum_{j=3}^{6} b_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

We saved the full model as `m2`. Assuming we saved the full model as `m3`, we can calculate the quantities need for the F test with as follows

```
> rss0 <- sum(m3$residuals^2)
> df0 <- m3$df.residual
> rss1 <- sum(m2$residuals^2)
> df1 <- m2$df.residual
> fstat <- (rss0-rss1)/(df0-df1)/(rss1/df1)
> fstat
[1] 5.973961
> 1-pf(fstat,df0-df1,df1)
[1] 0.001162835
```

The small p-value indicates that we evidence that age 18 height is related to non-height age 9 variables after controlling for age 9 height.

Of course, R has a command to make these tests easier for us. We can compute the two F-tests discussed so far using the `anova` command:

```
> anova(m2,m1)
Analysis of Variance Table

Model 1: HT18 ~ WT9 + HT9 + LG9 + ST9
Model 2: HT18 ~ WT2 + HT2 + WT9 + HT9 + LG9 + ST9
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     65 693.52
2     63 675.36  2    18.153 0.8467 0.4336

> anova(m3,m2)
Analysis of Variance Table

Model 1: HT18 ~ HT9
Model 2: HT18 ~ WT9 + HT9 + LG9 + ST9
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1     68 884.73
2     65 693.52  3    191.22 5.974 0.001163 **
```

One final exercise: when we print the summary of a model fit, the last line of the summary contains something like

```
F-statistic: 127.7 on 1 and 68 DF,  p-value: < 2.2e-16
```

You get a different F statistic for each of the three models we fit. Can you figure out what the reduced and full models are for these tests?

Chapter 6

Factors

Joseph Guinness - BTRY 6020

A factor is a qualitative variable that takes on one of several levels. Examples include dog breed, which might have levels greyhound, whippet, and Italian greyhound; religion, which could have many levels Christian, Jewish, Muslim, etc.; and experimental group, which might have four levels, control, treatment 1, treatment 2, and treatment 3. We usually treat the levels of the factor as exhaustive and mutually exclusive. For example, if we had levels Christian, Jewish, and Muslim, and some of the subjects identified as both Christian and Jewish, we would probably just create a fourth level, Christian + Jewish, for those subjects. We will use the generic letter J to refer to the number of levels of the factor.

When analyzing data that include factor variables, we sometimes want to know how the expected response–a quantitative variable–differs according to subjects' factor levels. For example, suppose we are conducting an experiment on drought tolerance of $J = 3$ different wheat varieties by planting each variety in four subplots, for a total of 12 subplots of wheat. The factor is variety, the three levels are the three different varieties that appear in the experiment, and the 12 responses are the yields from the 12 subplots.

We will study two different mathematical notations for data and models that use factors. One is simpler, and one is more complicated but crucial for making connections to multiple linear models.

The simple one first. Let y_{ij} be the ith response that has level j of the factor. In our wheat variety example, we would have the 12 responses

$$\underbrace{y_{11}, y_{21}, y_{31}, y_{41}}_{\text{variety 1}}, \underbrace{y_{12}, y_{22}, y_{32}, y_{42}}_{\text{variety 2}}, \underbrace{y_{13}, y_{23}, y_{33}, y_{43}}_{\text{variety 3}}$$

To investigate whether the different varieties have different expected yields, we want a model for y_{ij} that has a different expected yield for each variety j. This will do the trick:

$$Y_{ij} = b_0 + b_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

Under this model, the expected value of Y_{ij} is equal to $b_0 + b_j$, and the observed response is its expected value plus a normal error. We call this notation double subscript notation, owing to the two subscripts on y.

You might be asking why we don't simply write $Y_{ij} = b_j + \varepsilon_{ij}$. This would be simpler, but that's not how we do it. More on that later.

The other notation defines y_i as the ith response. So in our wheat variety example, we would have

$$\underbrace{y_1, y_2, y_3, y_4}_{\text{variety 1}}, \underbrace{y_5, y_6, y_7, y_8}_{\text{variety 2}}, \underbrace{y_9, y_{10}, y_{11}, y_{12}}_{\text{variety 3}}$$

We call this single subscript notation. Since the notation for the response does not contain any information about the level of the factor, we need an additional variable to link i to a level. This is achieved with dummy variables

$$x_{ij} = \begin{cases} 1 & \text{if the } i\text{th response has factor level } j \\ 0 & \text{if the } i\text{th response does not have factor level } j. \end{cases}$$

The model can then be written as

$$Y_i = b_0 + \sum_{j=1}^{J} b_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

This might seem a little strange because subject i can by definition possess only one level of the factor, but the right side of the equation contains information about all levels of the factor. The dummy variables make this work out nicely. In our wheat example, plot $i = 6$ has variety $j = 2$. The model for y_6 becomes

$$Y_6 = b_0 + b_1 x_{61} + b_2 x_{62} + b_3 x_{63} = b_0 + b_1 0 + b_2 1 + b_3 0 = b_0 + b_2.$$

As we hoped, the response from a subject with level 2 of the factor has expected value equal to $b_0 + b_2$, just as we had in double subscript notation. Now we see that double subscript and single subscript notation are two ways of writing down the same model.

Double subscript is simpler in that it avoids using dummy variables. Single subscript notation, however, has a clear connection to the multiple linear model because it has the exact same form as the multiple linear model. In this case, there are J covariates $x_{i1}, x_{i2}, \ldots, x_{iJ}$. This is incredibly powerful because it means that we can estimate factor models in the same way that we estimate

other multiple linear models, and we can conduct t-tests involving indidual comparisons of b_j variables, and we can conduct F-tests that involve multiple b_j variables.

The F-test will involve counting model degrees of freedom in factor models. The notation we use for this model makes the accounting slightly tricky, so forget about the notation for just a moment. The factor model says that responses from each of the J levels of the factor have a different expected value. If you wanted to know the mean part of the factor model, I would have to give you J numbers to indicate the J means, one for each level. No more, no fewer. Thus, the factor model has J model degrees of freedom. However, the notation tries to throw us off by giving us $J + 1$ parameters $b_0, b_1, \ldots, b_J$, so there are more parameters than model degrees of freedom. This is what I warned about–the model degrees of freedom is not always equal to the number of mean parameters.

When a model has more parameters than model degrees of freedom, we say that the model is overparmeterized. When a model is overparameterized, we have to say how we constrain the parameters when we estimate them. For example, we could impose the constraint $b_0 = 0$, leaving $b_1, \ldots, b_J$ to describe the factor model with J degrees of freedom. Or we could impose the constraint that $b_1 = 0$, or that $b_J = 0$, or perhaps the constraint that $b_1 + \cdots + b_J = 0$, which is more exotic, but a perfectly valid constraint.

Understanding constraints in factor models might be the single most confusing thing in this course, but you'll get used to it. In a perfect world, everyone would agree that we should impose the same constraint, say $b_0 = 0$, and then we could forget about this problem. However, that's not the case. Different software packages impose different constraints by default. For example, in R, the default constraint is $b_1 = 0$, whereas in SAS it's $b_J = 0$, and in JMP it's $b_1 + \cdots + b_J = 0$.

The scary part about all this–and the reason we spend time wrestling with this problem in this class–is that the interpretations of the individual parameters $b_0, \ldots, b_J$ change depending on what constraint has been imposed. Your software will impose a default constraint and spit out estimates of the parameters. In order to interpret them, you need to know what constraint has been imposed. One of the worst mistakes you can make in an analysis is to misinterpret the meaning of a parameter.

When figuring out the interpretation of an individual parameter under some constraint, remember that $b_0 + b_j$ is *always* interpreted as the expected value for responses that have factor level j. Before we impose a constraint, the individual parameters do not have interpretations. After a constraint is imposed, the individual parameters do have interpretations. To figure out the interpretations, our strategy is to look at combinations of expected values that isolate individual parameters. Remember, though, that even after the constraint is imposed, $b_0 + b_j$ is *always* interpreted as the mean for level j in the one-factor model.

Default R constraint $b_1 = 0$

We know that

$$b_0 + b_1 = \{ \text{ expected value for level 1 } \} .$$

This is always true, no matter the constraint. When we impose $b_1 = 0$, we have

$$b_0 + b_1 = \{ \text{ expected value for level 1 } \} = b_0 + 0 = b_0,$$

so we can interpret b_0 as the expected value (EV) for level 1. We also know that

$$(b_0 + b_2) - (b_0 + b_1) = \{ \text{ EV for level 2 minus EV for level 1 } \} = b_2$$

when $b_1 = 0$ is imposed. The parameter b_2 is interpreted as level 2 EV minus level 1 EV. We can go through the same argument to establish that under the default R constraint that $b_1 = 0$, b_j is interpreted as EV for level j minus EV for level 1.

Constraint $b_0 = 0$

This one is easier. Remember that

$$b_0 + b_j = \{ \text{ expected value for level } j \} = b_j \quad \text{when } b_0 = 0.$$

This means that b_j is interpreted as the mean for level j.

Default SAS constraint $b_J = 0$

Similar to the R constraint, b_0 is interpreted as the mean for level J, and b_j is interpreted as EV for level j minus EV for level J.

Default JMP constraint $b_1 + \cdots + b_J = 0$

This one is more tricky. Consider the average of the expected values,

$$\frac{1}{J}\sum_{j=1}^{J}(b_0 + b_j) = b_0 + \frac{1}{J}\sum_{j=1}^{J} b_j = b_0.$$

Therefore, b_0 is interpreted as the average of the expected values for the J levels. Since $b_0 + b_j$ is always the EV for level j, then

$$b_0 + b_j - b_0 = \{ \text{ EV for level } j \text{ minus average EV } \} = b_j,$$

which gives an interpretation for b_j.

If you remember one thing, remember that you have to be careful about this. If you remember two things, memorize the interpretations for the default R constraint.

You might be wondering why R imposes such a weird constraint $b_1 = 0$. Isn't $b_0 = 0$ a simpler constraint? In some ways it is, but there is a good reason why the R default is $b_1 = 0$. In many studies there is a control group in addition to one or several treatment groups. The quantity of interest is usually the EV

for the treatment group minus the EV for the control group. If we set the control group to be factor level 1, then b_0 is the EV for the control group, and $b_2, \ldots, b_J$ are EV for the treatment groups minus the EV for the control group. R will print out estimates and t-tests for $b_2, \ldots, b_J$, which is often what we are interested in learning from this analysis.

F-tests

Most commonly, we are interested in testing the hypothesis

$$H_0 : b_1 = b_2 = \cdots = b_J \quad \text{(Hypothesis A)},$$

which corresonds to a hypothesis that the EV is the same in every level. We call this Hypothesis A to distinguish it with another case, where we might be interested in whether all the treatment groups have the same EV. If level 1 is the control group, then this hypothesis is

$$H_0 : b_2 = \cdots = b_J \quad \text{(Hypothesis B)},$$

These hypotheses involve multiple parameters, and so we usually need an F-test. In the F-test the full model is (using single subscript notation)

$$\text{Full Model:} \quad Y_i = b_0 + \sum_{j=1}^{J} b_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

In Hypothesis A, where all the EVs are equal, the reduced model is

$$\text{Reduced Model:} \quad Y_i = b_0 + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

In Hypothesis B, where only treatment groups have equal EVs, the reduced model can be written as

$$\text{Reduced Model:} \quad Y_i = b_0 + b_1 x_{i1} + b_2 \sum_{j=2}^{J} x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

If that looks a little strange, take a minute, look back at the definition of the dummy variable, and convince yourself that the model corresponds to one with a two different EVs: $b_0 + b_1$ for level 1, and $b_0 + b_2$ for all the other levels.

To calculate the f statistic, we must fit both the full model and the reduced model, and save the residual sums of squares and the model degrees of freedom. For both hypotheses, the full model degrees of freedom is J, one EV for each level. In Hypothesis A, the reduced model degrees of freedom is 1, since we need only a single number to describe the common mean. This means that the f statistic is

$$f = \frac{(rss0 - rss1)/(J-1)}{rss1/(n-J)} \quad \text{(Hypothesis A } f \text{ statistic)},$$

where $rss0$ is the rss for the Hypothesis A reduced model. In Hypothesis B, the reduced model degrees of freedom is 2, one for the control group, and one for the treatment group. This means that the f statistic is

$$f = \frac{(rss0 - rss1)/(J-2)}{rss1/(n-J)} \quad \text{(Hypothesis B } f \text{ statistic)},$$

where $rss0$ is the rss for the Hypothesis B reduced model.

Hypotheses A and B are just two examples of things you might want to test for in factor models. We will see more examples in the lecture and the homework.

6.1 Marathon Data

We are interested in a picking a marathon where we can run the fastest time, so we decide to collect data on marathon performances to understand which marathons have the fastest course. There is a website called marathonguide.com that compiles results for nearly all (if not all) marathons. I wrote a python script for scaping results from the website and have created a dataset contain men's results from these marathons: For a result to be included in the dataset,

Marathon race	Course Label	Race Label
Boston Marathon 2018	1	1
Boston Marathon 2019	1	2
Chicago Marathon 2018	2	3
Chicago Marathon 2019	2	4
New York City Marathon 2018	3	5
New York City Marathon 2019	3	6

the person must have run at least 4 of the 6 races. I only considered exact matches of names and attempted to filter out erroneous matches by ensuring that individual people's ages did not differ by more than 2 years in the results.

We will perform several analyses of this dataset. The first one here, will not account for the runners' ability in the analysis; we will address that aspect later. Let $y_1, \ldots, y_n$ be the marathon time in minutes for the ith performance in the dataset. Let $j(i)$ be the course label associated with the ith performance, and let $k(i)$ be the race label associated with the ith performance. Consider the following model for the times

$$Y_i = b_0 + b_{j(i)} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

This model says the following It also says that an individual performances is equal to its expected value plus a normal error with variance σ^2, and the errors are independent. Take a moment to think about whether these assumptions are appropriate given what the data consist of.

When we fit the model in R, we get the following output

$b_0 + b_1$	expected time at Boston Marathon
$b_0 + b_2$	expected time at Chicago Marathon
$b_0 + b_3$	expected time at New York City Marathon

```
Call:
lm(formula = time_minutes ~ marathon, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-96.69 -36.12 -12.77  19.25 217.62

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      215.732      1.614 133.653  < 2e-16
marathonChicago Marathon           6.175      2.385   2.589  0.00969
marathonNew York City Marathon    11.707      2.376   4.927 8.87e-07

Residual standard error: 50.32 on 2623 degrees of freedom
Multiple R-squared:  0.00922,Adjusted R-squared:  0.008464
F-statistic:  12.2 on 2 and 2623 DF,  p-value: 5.301e-06
```

We see an intercept coefficient. This is the estimate for b_0. We also see coefficients for Chicago and New York City. These are estimates for b_2 and b_3. R has chosen to set $b_1 = 0$ (the Boston coefficient). Therefore, we interpret $\widehat{b}_0 = 215.732$ as the expected time in minutes at the Boston Marathon. $\widehat{b}_2 = 6.175$ means that we estimate that we expect a Chicago time to be 6.175 minutes slower than Boston, and $\widehat{b}_3 = 11.707$ means that we estimate that we expect a New York time to be 11.707 minutes slower than a Boston time. The standard errors on $\widehat{b}_2$ and $\widehat{b}_3$ tell us our uncertainty about the expected *difference* between a Boston time and either a Chicago or a New York time.

If we want standard errors for different comparisons, we can relevel the marathons by typing:

```
dat$marathon <- relevel(dat$marathon, "Chicago Marathon")
```

This makes Chicago Marathon the reference level. If we refit the model, we get the following table

```
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      215.732      1.614 133.653  < 2e-16
marathonBoston Marathon           -6.175      2.385  -2.589  0.00969
marathonNew York City Marathon     5.532      2.376   2.328  0.01997
```

Now, the comparisons are to the Chicago Marathon. Note how the estimates and the stadard errors change. Remember that the models don't change at all if we change the reference level; this is just a different choice for how we constrain the coefficients. In the second one, we set $b_2 = 0$ instead of $b_1 = 0$. You can

verify that nothing has changed by calculating $\widehat{b}_0 + \widehat{b}_j$ in both outputs.

We might also be interested in how marathon times differ in the two years. We can fit this model to probe that question:

$$Y_i = c_0 + c_{k(i)} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Recall that $k(i)$ is the label associated with each individual race, of which there are six. I've used c instead of b to help differentiate between the two models. If we fit this model in R, we get the following output

```
Call:
lm(formula = time_minutes ~ mar_year, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-95.99 -36.03 -13.12  19.43 215.53

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          217.822      2.319  93.944  < 2e-16
mar_yearBoston Marathon 2019          -4.053      3.230  -1.255  0.20956
mar_yearChicago Marathon 2018          6.140      3.362   1.826  0.06794
mar_yearChicago Marathon 2019          1.859      3.436   0.541  0.58848
mar_yearNew York City Marathon 2018   10.572      3.397   3.112  0.00188
mar_yearNew York City Marathon 2019    8.688      3.373   2.576  0.01005

Residual standard error: 50.32 on 2620 degrees of freedom
Multiple R-squared:  0.01048,Adjusted R-squared:  0.008597
F-statistic: 5.552 on 5 and 2620 DF,  p-value: 4.296e-05
```

Think about how you interpret these coefficients. What did R select as the reference level? In light of which level is the reference level, how do you interpret the other coefficients and their standard errors?

It seems that there might be some evidence that the times differ within a marathon course, between the two years. To formally test this hypothesis, we should do an F test to compare the first model (course only) as the reduced model and the second model (course and year) as the full model. The reduced model as 3 model degrees of freedom, and the full model has 6 model degrees of freedom. The reduced model is a special case of the full model. To see why, we get the reduced model by imposing the following hypothesis on the full model:

$$c_1 = c_2 \text{ and } c_3 = c_4 \text{ and } c_5 = c_6.$$

If that hypothesis is true, then the expected times in the two years is the same for each marathon, but can differ among marathons, which is precisely the assumption in the reduced model.

We can use the `anova` command to do the F test:

```
> anova(m1,m2)
Analysis of Variance Table

Model 1: time_minutes ~ marathon
Model 2: time_minutes ~ mar_year
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1   2623 6642619
2   2620 6634135  3    8483.4 1.1168 0.3409
```

This says that we don't have strong evidence that the times differ in the two years within a marathon course. More to come on this dataset.

Chapter 7

Factors and Numeric Covariates

Joseph Guinness - BTRY 6020

So far, we have seen models with numeric covariates, and we've seen separate models with factor covariates. It's probably not surprising that you can have both factor covariates and numeric covariates together in the same model. That is the topic of this chapter. We will also talk about the concept of an additive model and see our first example of a non-additive model, also known as an interaction model.

Suppose that we have one factor with J levels and, for simplicity's sake, a single numeric covariate. We have n subjects and observe

$$\begin{aligned} \text{responses}: &\quad y_1, \ldots, y_n \\ \text{numeric covariate}: &\quad x_i, \ldots, x_n. \end{aligned}$$

We also know the factor level that each of the subjects belongs to. Since we are using x for the numeric covariate, we make a slight change from the previous chapter and use z for the factor dummy variables:

$$z_{ij} = \begin{cases} 1 & \text{if subject } i \text{ has factor level } j \\ 0 & \text{if subject } i \text{ does not have factor level } j. \end{cases}$$

Using this notation, we can write the factor-numeric additive model for y_i as

$$Y_i = a_0 + \sum_{j=1}^{J} a_j z_{ij} + b_0 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

To see how this model works, suppose that subject 7 has factor level 3. Then the model simplifies to

$$Y_7 = a_0 + a_3 + b_0 x_7 + \varepsilon_7 = \Big(a_0 + a_3\Big) + b_0 x_7 + \varepsilon_7.$$

In this example, the relationship between $E(Y_7)$ and x_7 is linear as a function of x_7, with intercept $a_0 + a_3$ and slope b_0.

In general, if subject i has factor level j, $E(Y_i)$ is a linear function of x_i with intercept $a_0 + a_j$ and slope b_0. This model says that each factor level gets its own intercept $a_0 + a_j$, but all factor levels share a common slope, b_0. Therefore the model degrees of freedom is $J + 1$, accounting for the J intercepts and 1 slope.

In contrast, the model specifies $J + 2$ parameters, $a_0, a_1, \ldots, a_j, b_0$, and so the model is overparameterized. Just as in the factor-only model, your software will impose some constraint when estimating the model parameters. In R, the default sets $a_1 = 0$, which means that a_0 is interpreted as the intercept for level 1, and a_j is interpreted as the difference between the level j intercept and the level 1 intercept. This is analogous to how R treats the parameters in the factor-only model.

The factor-numeric model above is an example of an **additive model**. Additivity is a very important concept that we will focus our attention on in the next few weeks. We give a definition here.

Definition 3 *A model with multiple covariates is **additive** if the effect of each covariate does not depend on the value of the other covariates.*

To demonstrate why our factor-numeric model is additive, we first need to say what we mean when we talk about the effect of a variable. More precisely, the effect of a variable is the impact that changing the variable has on the expected response.

Let's see how this works in our factor-numeric model by looking at some ficticious height and sex data. The response is height at age 18, the numeric covariate is height at age 9, and the factor is sex, with two levels in this dataset, male (level 1) and female (level 2).

i	HT9	Sex	HT18
1	133	male	180
2	135	male	185
3	133	female	170
4	135	female	172

To see the effect of changing the numeric covariate while holding the factor level constant, we can compare subjects 1 and 2, which are are both male. Under our factor-numeric model

$$E(Y_2) - E(Y_1) = \Big(a_0 + a_1 + b_0 x_2\Big) - \Big(a_0 + a_1 + b_0 x_2\Big) = b_0\Big(x_2 - x_1\Big).$$

The **effect** of changing the covariate from x_1 to x_2 is simply the slope b_0 multiplied by the change in the covariate. Likewise, we could compare subjects 3

and 4, which are both female:

$$E(Y_4) - E(Y_3) = \Big(a_0 + a_2 + b_0 x_4\Big) - \Big(a_0 + a_2 + b_0 x_3\Big) = b_0\Big(x_4 - x_3\Big).$$

Again, the effect is to increase the expected response by b_0 times the change in the numeric covariate. *Thus, the effect of changing the numeric covariate does not depend on the level of the factor covariate.*

To demonstrate that the model is additive, we also need to check that the effect of changing the factor level does not depend on the value of the numeric covariate. To check this, we can compare subjects 1 and 3 or subjects 2 and 4. Both pairs have the same height at age 9 but differ in their sex. We find that

$$\begin{aligned} E(Y_3) - E(Y_1) &= \Big(a_0 + a_2 + b_0 x_3\Big) - \Big(a_0 + a_1 + b_0 x_1\Big) \\ &= a_2 - a_1 + b_0\Big(x_3 - x_1\Big) = a_2 - a_1 \\ E(Y_4) - E(Y_2) &= \Big(a_0 + a_2 + b_0 x_4\Big) - \Big(a_0 + a_1 + b_0 x_2\Big) \\ &= a_2 - a_1 + b_0\Big(x_4 - x_2\Big) = a_2 - a_1 \end{aligned}$$

The effects are the same, $a_2 - a_1$! This means that *the effect of changing the factor level did not depend on the value of the numeric covariate.* We got the same effect when the numeric covariate was 133 as we did when the numeric covariate was 135.

Taken together, this means that the factor-numeric model is an additive model; the effect of changing the numeric covariate did not depend on the level of the factor, and the effect of changing the factor level did not depend on the value of the numeric covariate.

The symbols, subscripts, and comparisons can be a bit overwhelming, so a simple plot might demonstrate the concept of additivity better. In Figure 7.1, we plot the Berkeley guidance study data and add lines for an estimated additive model. We can see that the slopes are equal but the two lines have different intercepts. Parallel slopes are a definitive feature of additive factor-numeric models.

Additivity might seem like an inevitable property of multiple linear models. But it is not. Consider the following model

$$Y_i = \Big(a_0 + \sum_{j=1}^{J} a_j z_{ij}\Big) + \Big(b_0 x_i + \sum_{j=1}^{J} b_j z_{ij} x_i\Big) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

To analyze this model, suppose again that subject 7 has factor level 3. Then the model simplifies to

$$Y_7 = a_0 + a_3 + b_0 x_i + b_3 x_i + \varepsilon_i = \Big(a_0 + a_3\Big) + \Big(b_0 + b_3\Big) x_i + \varepsilon_i.$$

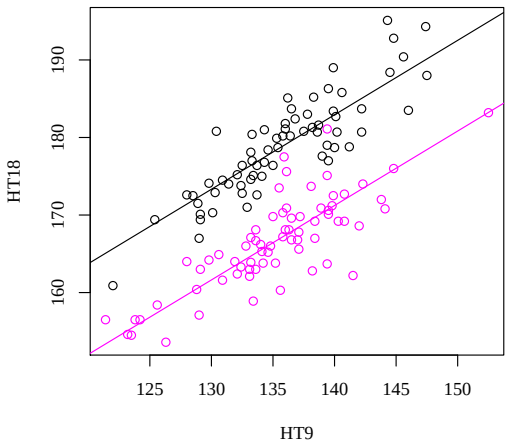


Figure 7.1: Berkeley guidance study data, girls in magenta, boys in black. Estimated additive model lines drawn.

Once we group the terms, we can see that this model says that the expected response is a linear function of the numeric covariate, with intercept $a_0 + a_3$ and slope $b_0 + b_3$, both of which depend on the factor level. This means that this model has $2J$ model degrees of freedom; we need two numbers, an intercept and a slope, per factor level to describe the mean part of the model.

As before, let's analyze the model by considering comparisons from our ficticious dataset. To see the effect of changing the numeric covariate while holding the factor level constant, we need to compare subjects 1 and 2 (both male but different heights at age 9) and subjects 3 and 4 (both female but different heights at age 9). We get

$$E(Y_2) - E(Y_1) = \Big(b_0 + b_1\Big)\Big(x_2 - x_1\Big)$$
$$E(Y_4) - E(Y_3) = \Big(b_0 + b_2\Big)\Big(x_4 - x_3\Big).$$

These are different effects! The effect for the male subjects is $b_0 + b_1$ times the change in the numeric covariate, whereas the effect for the female subjects is $b_0 + b_2$ times the change in the numeric covariate.

Likewise, we can consider the effect of changing the factor level while holding the numeric covariate constant. This is achieved by comparing subjects 1 and

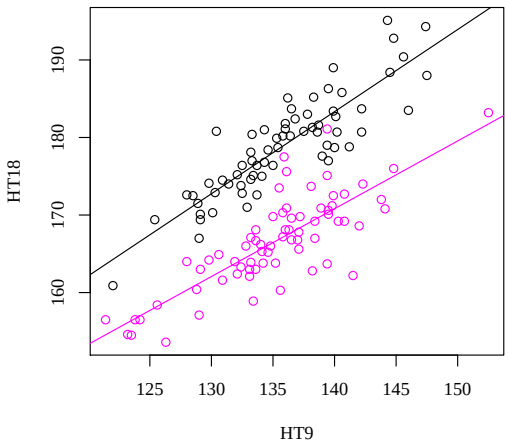


Figure 7.2: Berkeley guidance study data, girls in magenta, boys in black. Estimated interaction model lines drawn.

3 and subjects 2 and 4.

$$\begin{aligned} E(Y_3) - E(Y_1) &= \Big(a_0 + a_2 + b_0 x_3 + b_2 x_3\Big) - \Big(a_0 + a_1 + b_0 x_1 + b_1 x_1\Big) \\ &= a_2 - a_1 + 133(b_2 - b_1) \\ E(Y_4) - E(Y_2) &= \Big(a_0 + a_2 + b_0 x_4 + b_2 x_4\Big) - \Big(a_0 + a_1 + b_0 x_2 + b_1 x_2\Big) \\ &= a_2 - a_1 + 135(b_2 - b_1) \end{aligned}$$

These two comparisons are also different! We conclude that this model is not additive.

There is a more familiar name for models that are not additive. We typically refer to them as interaction models. Interaction models are commonly understood as models in which the covariates are multiplied together. This is true, but knowing that the covariates are multiplied together doesn't help us understand what the model is assuming.

It's best to view an interaction model as a non-additive model. **In interaction models, the effect of one covariate may depend on the value of the other covariates.**

In Figure 7.2, we plot the same data, but with the fitted interaction model lines plotted. The hallmark difference is that the lines are no longer parallel. The effect of increasing age 9 height is larger for boys than it is for girls. Likewise, the effect of changing sex is smaller at the low end of age 9 height than it is at

the high end of age 9 height.

It is common to test for the presence of an interaction. For this test, the full model is the factor-numeric interaction model, and the reduced model is the additive factor-numeric model. The numerator of the F statistic has $2J - (J + 1) = J - 1$ degrees of freedom, and the denominator has $n - 2J$ degrees of freedom.

Chapter 8

Models with Multiple Factors

Joseph Guinness - BTRY 6020

Just like we can have models with multiple numeric covariates, we can have models with multiple factors. This chapter will show how to specify these models and how to understand them in terms of additivity and non-additivity.

Suppose we have two factors. The first factor has J levels, and the second factor has K levels. Figure 8.1 shows one example. The response is the logarithm of battery life in hours, and the two factors are battery type, which has 3 levels, and the temperature of the experiment, which also has 3 levels. Temperature could be viewed as a numeric variable, but we will treat it as a factor with three possible levels.

To introduce the model, we will use notation with multiple subscripts. Later on, we will show single subscript notation. Define y_{ijk} to be the ith response that had factor 1 level j and factor 2 level k. For example, the battery data responses might be listed as

$$y_{111}, y_{211}, y_{311}, y_{411}, y_{121}, y_{221}, y_{321}, y_{421}, y_{131}, \ldots,$$

which means that the first four observations had battery type 1 and temperature 15 degrees (levels 1 and 1), the next four had battery type 2 and temperature 15 degrees (levels 1 and 2), and so on.

The additive two factor model for y_{ijk} is

$$Y_{ijk} = b_0 + b_j + c_k + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2).$$

The easiest way to think about the model is to write down what the means are for the various combinations of levels. This is shown in Table 8.5.

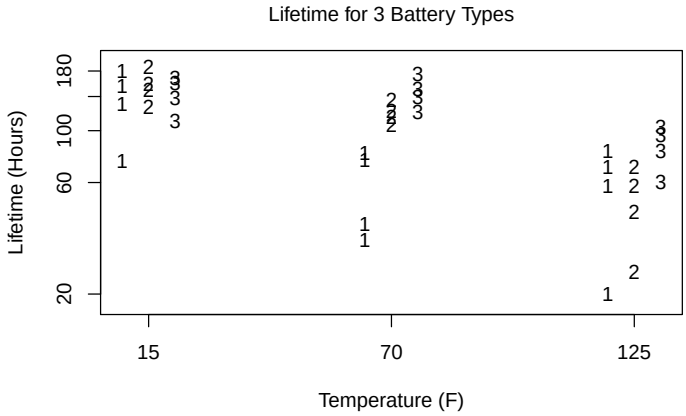


Figure 8.1: Battery Data.

	$k=1$	$k=2$	$k=3$
$j=1$	$b_0+b_1+c_1$	$b_0+b_1+c_2$	$b_0+b_1+c_3$
$j=2$	$b_0+b_2+c_1$	$b_0+b_2+c_2$	$b_0+b_2+c_3$
$j=3$	$b_0+b_3+c_1$	$b_0+b_3+c_2$	$b_0+b_3+c_3$

Table 8.1: Expected valules for a two factor additive model with 3 levels for both factors. No overparameterization constraints have been applied.

We are calling this an additive two factor model. Recall that in an additive model, the effect of one variable does not depend on the value of the other variable. In the context of factor variables, this means that the effect of changing one of the factor levels does not depend on the level of the other factor. We can see that our model is additive by comparing means. For example, the effect of changing factor 1 from level 1 to level 2 is $b_2 - b_1$ regardless of the level of factor 2 (recall that the middle subscript refers to the level of factor 1 and the third subscript refers to the level of factor 2):

$$E(Y_{i21}) - E(Y_{i11}) = (b_0 + b_2 + c_1) - (b_0 + b_1 + c_1) = b_2 - b_1$$
$$E(Y_{i22}) - E(Y_{i12}) = (b_0 + b_2 + c_2) - (b_0 + b_1 + c_2) = b_2 - b_1$$
$$E(Y_{i23}) - E(Y_{i13}) = (b_0 + b_2 + c_3) - (b_0 + b_1 + c_3) = b_2 - b_1$$

In table form, these differences arise by comparing row 2 to row 1 of Table 8.5. We would find similar results if we changed the level of factor 1 from level 1 to level 3 ($b_3 - b_1$) or from level 2 to level 3 ($b_3 - b_2$).

Likewise, the effect of changing the level of factor 2 does not depend on the level of factor 1. For example, suppose we change the level of factor 2 from level 2 to level 3. The difference in expected value is always $c_3 - c_2$:

$$E(Y_{i13}) - E(Y_{i12}) = (b_0 + b_1 + c_3) - (b_0 + b_1 + c_2) = c_3 - c_2$$
$$E(Y_{i23}) - E(Y_{i22}) = (b_0 + b_2 + c_3) - (b_0 + b_2 + c_2) = c_3 - c_2$$
$$E(Y_{i33}) - E(Y_{i33}) = (b_0 + b_3 + c_3) - (b_0 + b_3 + c_2) = c_3 - c_2.$$

These differences arise by comparing column 3 to column 2 in Table 8.5.

Counting the model degrees of freedom in the two factor model is a bit tricky, but not out of our reach. The first thing to note is that the model cannot have more than JK (J times K) model degrees of freedom. This is because there are JK possible combinations of the two factor levels, and so there are only JK possible expected values.

The calculations we did above indicate that, in the additive model, these JK expected values have to follow a specific pattern. For example, the difference between (row 2, column 1) and (row 1, column 1) must be the same as the difference between (row 2, column 2) and (row 1, column 2). This restriction imposed on the expectations implies that the model degrees of freedom is less than JK.

To gain some intuition for what the actual model degrees of freedom is, let's think about how many numbers you would need to provide in order for me to know what all of the expected values are. And let's start with the simplest example, both factors have two levels. Table 8.2 shows a table with three of the expected values filled in. Can you see why these three numbers allow you to fill in the fourth expected value?

	$k = 1$	$k = 2$
$j = 1$	11	15
$j = 2$	14	??

Table 8.2:

You can fill in the fourth because you know that 14 is 3 more than 11, so the missing entry must be 3 more than 15. Likewise, you know that 15 is 4 more than 11, and so the missing value must be 4 more than 14. Both ways you get 18 for the missing value. If you took away any of these numbers, you would not be able to fill in the rest of the table, and so the model degrees of freedom here has to be 3. It's the minimum number of numbers I need in order to completely fill in the table of expected values.

Let's try the next most complicated example, factor 1 has 2 levels, and factor 2 has three levels.

	$k=1$	$k=2$	$k=3$
$j=1$	9	12	4
$j=2$	8	??	??

Table 8.3:

	$k=1$	$k=2$	$k=3$
$j=1$	b_0	b_0+c_2	b_0+c_3
$j=2$	b_0+b_2	$b_0+b_2+c_2$	$b_0+b_2+c_3$
$j=3$	b_0+b_3	$b_0+b_3+c_2$	$b_0+b_3+c_3$

Table 8.4: Expected valules for a two factor additive model with 3 levels for both factors. Default R overparameterization constraint has been applied.

Since the difference between the second row and the first row has to be the same in each column, we can easily fill in the two missing values as 11 and 3. Furthermore, you need all of the supplied information; if I had left out the 4 ($j=1$, $k=3$), you would not have been able to fill in the second missing value. This model must have 4 model degrees of freedom.

In the additive model, you need to have the entire first row and the entire first column (and nothing less) in order to fill in the entire table. This means that the additive two factor model has $J+K-1$ model degrees of freedom, arising from the fact that the first column has J entries, the first row has K entries, and the $(1,1)$ entry got counted twice, so we subtract 1.

The total number of parameters in the additive two factor model is $J+K+1$, which is two larger than the model degrees of freedom, so we need to specify two constraints when fitting the model. In R, the default constraint is to set $b_1=0$ and $c_1=0$. Under this constraint, we can simplify the table of expected values for out battery example:

By now, you are a pro at interpreting the parameters after a constraint has been applied. Here, b_0 is the expected value for levels $(1,1)$, b_j is the expected value of factor 1 level j minus expected value of factor 1 level 1, and c_j is the expected value of factor 2 level j minimus expected value of factor 2 level 1.

The two factor interaction model places no restrictions on how the various entries of the table of expected values are related. This means that we can specify them willy nilly, giving us JK model degrees of freedom.

	$k=1$	$k=2$	$k=3$
$j=1$	b_0	b_0+c_2	b_0+c_3
$j=2$	b_0+b_2	$b_0+b_2+c_2+(bc)_{22}$	$b_0+b_2+c_3+(bc)_{23}$
$j=3$	b_0+b_3	$b_0+b_3+c_2+(bc)_{32}$	$b_0+b_3+c_3+(bc)_{33}$

Table 8.5: Expected valules for a two factor interaction model with 3 levels for both factors. Default R overparameterization constraint has been applied.

Our notation for the two factor interaction model is

$$Y_{ijk} = b_0 + b_j + c_k + (bc)_{jk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0,\sigma^2).$$

The notation $(bc)_{jk}$ is a common one. The symbol should be interpreted as a single parameter (not a product of parameters), identified by the two subjects j and k. If we had two levels for each factor, we would have four parameters $(bc)_{11}$, $(bc)_{12}$, $(bc)_{21}$, and $(bc)_{22}$.

This model is REALLY overparameterized. There are $JK+J+K+1$ parameters but only JK model degrees of freedom. The default R constraints set any parameter with a "1" subscript to zero. Applying this constraint yields the following table of expected values for our battery example:

The interpretation of individual parameters gets a little complicated now, but the interpretations do exist. The interaction parameters get interpereted as "differences of differences." To see this, consider the difference between the difference between row 2 and row 1 for columns 1 and 2:

$$\begin{aligned}\Big[E(Y_{i22}) - E(Y_{i12})\Big] - \Big[E(Y_{i21}) - E(Y_{i11})\Big] \\ &= \Big[\Big(b_0+b_2+c_2+(bc)_{22}\Big) - \Big(b_0+c_2\Big)\Big] - \Big[\Big(b_0+b_2\Big) - \Big(b_0\Big)\Big] \\ &= (bc)_{22}.\end{aligned}$$

The interaction parameter $(bc)_{22}$ is determined as a difference of differences. Recall that in the additive model, these differences are the same, and so the difference of the differences is zero. So a t-test for $(bc)_{22} = 0$ is testing for whether there is an interaction between the first two levels of both factors.

The two factor models can be written in multiple linear regression (single subscript) format. Let x_{ij} be a dummy variable for the jth level of factor 1, and let z_{ik} be a dummy variable for the kth level of factor 2. The additive two factor model is

$$Y_i = b_0 + \sum_{j=1}^{J} b_j x_{ij} + \sum_{k=1}^{K} c_k z_{ik} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2).$$

The interaction model is

$$Y_i = b_0 + \sum_{j=1}^{J} b_j x_{ij} + \sum_{k=1}^{K} c_k z_{ik} + \sum_{j=1}^{J}\sum_{k=1}^{K} (bc)_{jk} x_{ij} z_{ik} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Chapter 9

Numeric-Numeric Interactions and Quadratic Models

Joseph Guinness - BTRY 6020

This chapter covers interpretation of interactions between numeric covariates. We also discuss models that are quadratic–rather than linear–in the numeric covariate. Much of this chapter is an exercise in grouping terms in the model equation for the purpose of helping us conceptualize the various models.

9.1 Review: the additive multiple linear model

Suppose we have responses $y_1, \ldots, y_n$ and two numeric covariates $x_{11}, \ldots, x_{n1}$ and $x_{12}, \ldots, x_{n2}$. The additive multiple linear model for y_i is

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Why do we say that this model is addtive? The effect of x_{i1} does not depend on the value of x_{i2} (and vice versa). To see why, consider grouping the terms in the model as

$$Y_i = \Big(b_0 + b_2 x_{i2}\Big) + b_1 x_{i1} + \varepsilon_i,$$

which shows us that for a given value of x_{i2}, $E(Y_i)$ is a linear function of x_{i1}. The intercept is $b_0 + b_2 x_{i2}$, which does depend on x_{i2}, but the slope–also known as the effect of x_{i1}–is b_1, which does not depend on the value x_{i2}. We could

rearrange the model analogously as

$$Y_i = \Big(b_0 + b_1 x_{i1}\Big) + b_2 x_{i2} + \varepsilon_i,$$

to see that the effect of x_{i2} does not depend on the value of x_{i1}. The intercepts change, but the slopes do not. Hence, the model is additive in x_{i1} and x_{i2}.

9.2 Numeric Interaction Model

If we wish to estimate a model in which the effects of each variable depend on the value of the other variables, we need an interaction model. The simplest interaction model is

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_{12} x_{i1} x_{i2} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The first thing to note is that, even though it might not be immediately obvious, this model does fit into our multiple linear model framework. Define $x_{i3} = x_{i1} x_{i2}$. Then the model can be rewritten as

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

which is the same model but with different symbols. So we can use the same techniques we have learned throughout the semester to estimate the model and make inferences about it–least squares, t-tests, F-tests.

To see why this is an interaction model, let's rewrite the model in two different ways:

$$\text{First Representation:} \quad Y_i = \Big(b_0 + b_1 x_{i1}\Big) + \Big(b_2 + b_{12} x_{i1}\Big) x_{i2} + \varepsilon_i,$$

$$\text{Second Representation:} \quad Y_i = \Big(b_0 + b_2 x_{i2}\Big) + \Big(b_1 + b_{12} x_{i2}\Big) x_{i1} + \varepsilon_i.$$

The first representation is linear in x_{i2}, and the second is linear in x_{i1}. In both representations, both the intercept and the slope depend on the other variable. Hence, this is an interaction model, that is, not an additive model.

In this model, the intercepts and slopes depend on on the other variable in a very specific way. For example, in the first representation (linear in x_{i2}), the intercept and slope are

$$\text{intercept: } b_0 + b_1 x_{i1}$$
$$\text{slope: } b_2 + b_{12} x_{i1}.$$

When the model is viewed as linear in x_{i2} (the first representation), both the intercept and the slope are linear functions of x_{i1}.

We could go through the same exercise for the second representation (linear in x_{i1}). We would see that both the intercept and slope are linear in x_{i2}. In particular the slope is $b_1 + b_{12} x_{i2}$. Note that the "slope of the slope" is b_{12} in both representations.

9.3 Quadratic Models

We have focused a lot on linear models, but we don't really think that everything is linear. Sometimes we want a curved model. The simplest such model is a quadratic model, for example this one in terms of a single covariate x_{i1}:

$$Y_i = b_0 + b_1 x_{i1} + b_{11} x_{i1}^2 + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

We can usefully conceptualize a quadratic model as a model in which x_{i1} interacts with itself. To see this, consider the alternative but equivalent representation

$$Y_i = b_0 + \Big(b_1 + b_{11} x_{i1}\Big) x_{i1} + \varepsilon_i.$$

Think of this as a linear model in terms of x_{i1} in which the slope depends on the value of x_{i1}, that is, the slope is equal to $b_1 + b_{11} x_{i1}$. It is a model where the effect of x_{i1} depends on where along the number line x_{i1} is.

We can also have quadratic models in two or more covariates. For example

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_{12} x_{i1} x_{i2} + b_{11} x_{i1}^2 + b_{22} x_{i2}^2 + \varepsilon_i.$$

It should be no surprise that this model can be rewritten either as a quadratic model in x_{i1} or as a quadratic model in x_{i2} with coefficients that depend on either x_{i2} or x_{i1}.

$$Y_i = \Big(b_0 + b_2 x_{i2} + b_{22} x_{i2}^2\Big) + \Big(b_1 + b_{12} x_{i2}\Big) x_{i1} + b_{11} x_{i1}^2 + \varepsilon_i$$
$$Y_i = \Big(b_0 + b_1 x_{i1} + b_{11} x_{i1}^2\Big) + \Big(b_2 + b_{12} x_{i1}\Big) x_{i2} + b_{22} x_{i2}^2 + \varepsilon_i.$$

We can see that the two-variable quadratic model can be viewed as a interaction model, because the effect of x_{i1} depends on the value of x_{i2}. In particular, the coefficient multiplying x_{i1} depends on x_{i2}.

Quadratic models are particularly useful when the goal of the analysis is to find the values of the covariate that maximize the expected response.

9.4 Beyond quadratics: polynomial models

Sometimes even a quadratic model is not flexible enough to capture the dependence of the response on the coviarate. Linear and quadratic models are special cases of a wider class of models called polynomial models. A polynomial model of degree d is

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i1}^2 + \cdots b_d x_{i1}^d + \varepsilon_i = \sum_{j=0}^{d} b_j x_{i1}^j + \varepsilon_i.$$

Sometimes the polynomial model can be useful when the covariate is a factor whose levels are ordered. Suppose the covariate is a response to a survey question whose possible answers are never, rarely, sometimes, often, or always. This factor variable has five levels, but the levels possess a relative structure: often is greater than sometimes, and always is greater than often, and so on. We could assign the numbers 1 to 5 to the responses, and fit a polynomial model of degree 4, as in

$$Y_i = \sum_{j=0}^{4} b_j x_{i1}^j + \varepsilon_i.$$

If you fit this model and compared it to the fit from the factor model, you would find that the two models have exactly the same residual sum of squares! This is because any set of 5 means can be represented by a polynomial of degree four. In simpler terms, any two points can be captured by a line, and any three points can be captured by a quadratic, and so on.

In our ordered factor example, we might think that having a model with 5 model degrees of freedom is too much flexibility, and so we can consider fitting polynomials of degree less than 4. It might be the case that a quadratic model is flexible enough for our purposes. We can compare the models with F-tests because the quadratic model is a special case of a polynomial of degree 4 ($b_3 = 0$ and $b_4 = 0$).

Chapter 10

Principles of Experimental Design

Joseph Guinness - BTRY 6020

Up to now, we have discussed appropriate ways of analyzing data after it has been collected. As a researcher, you will usually have control over how to analyze your data, and in many cases also have control over how to collect the data in the first place. The manner in which an experiment is performed can have substantial impact on our ability to draw conclusions from the data, so we must think carefully about the decisions we make when designing an experiment, which includes thinking about how we plan to model and analyze the data after it is collected.

Discussing experimental design requires some new terminology:

Experimental Material: The collection of physical material that you plan to measure and collect data from.

Treatment: Condition that can be applied to experimental material and controlled by the experimenter.

Partition: A division of the experimental material into groups wherein each piece of material belongs to exactly one group.

Unit Partition: The finest possible partition in which no two groups necessarily receive the same treatment.

Experimental Units: The groups in the unit partition.

Experimental Design: An assignment of treatments to experimental units.

Response: Measured outcome of interest.

To make some sense out of these definitions, let's consider a couple of examples. In the first, suppose that you have an agricultural field, and you want to study the effect of a new type of fertilizer. Your question is whether the new fertilizer produces higher yields than the fertilizer you used last year. The fertilizer will be applied with a spreader attached to the back of a tractor. It's a real pain to switch the fertilizers, so you decide that you have to apply the same treatment to all locations within the same row of the field; after you're done spreading in each row, you can either spread the same fertilizer again in the next row or switch fertilizers. You decide to switch fertilizer after every row, resulting in a pattern of alternating old and new fertilizers. At the end of the growing season, you will measure the total yield within each row.

Experimental Material: the field

Tratment: fertilizer, factor with two levels: "old" and "new"

Partition: division of field into subfields

Unit Partition: division of the field into rows, because every location within a row must receive the same treatment (fertilizer), but different rows may receive different treatments.

Experimental Units: rows

Experimental Design: alternate old and new fertilizer from one row to the next

Response: yield

In the second example, we want to know how a new hormone treatment affects the birth weight of rat pups. In the experiment, we inject either the hormone or a placebo into female rats, impregnate them, and then measure the birth weights of all the individual rat pups after they are born. You decide to randomly assign the hormone or placebo to each female rat. You will evaluate the hormones by measuring the birth weight of each individual rat pup.

Experimental Material: female rats

Tratment: injection, either hormone or placebo

Partition: division of female rats into groups

Unit Partition: one female rat in each group, because we are allowed to give either hormone or placebo to each female rat.

Experimental Units: individual female rats

Experimental Design: random assignment

Response: weight of individual rat pups. Since a female rat may have more than one rat pup, there may be more individual responses than there are experimental units.

Aside from the treatments, there may be other variables of interest that we want to account for. In the rat example, we might also keep track of the weight of the female rats, and which male rat impregnated each female rat. We can include the weight as a numeric covariate and the male rat as a factor covariate when we model the data to account for their effects. We probably also want to think about how to assign the treatments in light of the differing fathers and differing weights of the mothers. We might even conceptualize the male rat as another treatment, since we have control over this aspect of the experiment.

The goal of an experimental design is to give ourselves the best chance of answering our question of interest. Before formalizing some ways of defining what it means to give ourselves the "best chance", let's look at some bad designs and discuss why they are bad. Suppose we have 20 female rats in the rat hormone example.

Bad Design 1: Assign 18 female rats to the placebo and 2 to the hormone treatment. This doesn't give us much information about the hormone.

Bad Design 2: Assign 18 female rats to the hormone group and 2 to the placebo group. This doesn't give us much information about the control group, which in turn doesn't give us much information about the *difference* between the control group and the treatment group.

Bad Design 3: Randomly assign treatments to the female rats, but impregnate the 10 female rats who got the placebo with male rat # 1, and impregnate the 10 female rats who got the hormone with male rat # 2. In this design, we would not be able to distinguish between the effect of the treatment and the effect of the male rat because they are perfectly confounded.

Bad Design 4: Assign the 10 lightest female rats to the control group and the 10 heaviest rats to the hormone group. This might be ok if weight of female rat has little effect on weight of rat pups (but weight probably does have an effect). If we think that there is an interaction between weight and the treatment, we'll have a difficult time distinguishing between the effect of weight and the effect of the hormone.

These may or may not seem like obviously bad ideas to you. Good designs are often quite intuitive. In my experience, researchers are adept at selecting good designs when they have sufficient control over the experiment, even when they have little or no training in experimental design or statistics. The problems arise more often when some aspect of the experiment that is outside of their control places a restriction on how the experiment can be conducted, and the researchers fail to fully appreciate how the restriction on the experiment impedes their ability to distinguish between the effects of the treatment and the confounding variables.

For example, suppose there are two different hormones of interest. One of the hormones is on backorder from the supplier. The researcher has limited time to do the experiment, so she obtains the other hormone and ten female rats from

the animal lab and conducts the experiment on 10 female rats using the first hormone while waiting for the backordered second hormone. When the second hormone finally arrives, she goes and gets ten more female rats from the animal lab, and conducts the second half of the experiment using the second hormone. However, in the meantime, the animal lab has changed their supplier, and the researcher notices that the second group of female rats is heavier on average. In the back of her mind, she knows that this might impact the weights of the offspring, but she convinces herself that this effect should be small relative to the effect of the hormone treatment.

Let us now introduce two key concepts of experimental design, **randomization** and **blocking**. Inevitably, asside from the treatments, the units will have other attributes that influence the response. Randomization and blocking help us avoid unequal allocation of treatments to units with respect to these other attributes. Randomization generally refers to a random allocation of treatments to units. For example, if we are assigning hormone or placebo to 20 female rats, for each female rat, we might flip a coin; if the coin lands heads, we assign it the placebo, tails the hormone. Alternatively, we might assign each female rat a number between 1 and 20, then place 20 numbered pieces of paper into a hat, then shuffle the numbers and draw 10 of the numbers. The hormone treatment is given to rats with the drawn numbers, the placebo to the others.

Randomization protects us against a grossly unequal allocation of treatments with respect to other attributes. However, it is not a foolproof strategy. By random chance, we might accidentally assign the hormone treatment to the lightest female rats. Blocking is a strategy for deliberately avoiding such situations by dividing the experimental units into groups that are similar with respect to an attribute that we expect to influence the response, and then making sure that the treatments are assigned equally with respect to the attribute. We will study block designs more thoroughly in a later chapter, but consider this simple example. If we expect weight of the female rat to influence the weight of the rat pups, we can group the 20 female rats into 10 groups, each group containing two female rats of similar weight. Then to assign the treatments, we randomly assign one from each group to the hormone and one to the placebo.

Of course, we cannot use blocking to avoid unequal allocation over unobservable attributes. The general strategy is to block to block over observable attributes and randomize to avoid grossly unequal allocation over all unobservable attributes.

As mentioned above, the goal of an experimental design is to give ourselves the best chance of answering our question of interest. This can be formalized with the concept of power introduced in Chapter 1. Suppose that we have a parameter of interest, b_1, and answering our question of interest corresponds to testing the hypothesis:

$$H_0 : b_1 = 0.$$

Given a particular experimental design D, a statistical model for the data, a decision rule, and a hypothesized "true" value of $b_1 = b_1^*$, we can calculate the power as

$$\text{Power}(b_1^*, D) = P(\text{ reject } H_0 \text{ when } b_1 = b_1^*).$$

This notation is slightly different than before because we have included the design D as something we can vary in the power calculation. For a given parameter value b_1^*, our goal is to pick the design D that maximizes the power, since we can get a solid answer to our question of interest when we reject the hypothesis.

It probably seems like it would be pretty complicated to calculate the power for every design, and sometimes it is, but in many cases the answers are simple and intuitive. One simplifying feature is that often the design that maximizes the power is the design that minimizes the standard error of the parameter estimate. So instead of doing a complicated power calculation, we can simply calculate the variance of the estimator and select a design that minimizes the variance.

Suppose that we have 20 female rats, and consider the oversimplistic scenario: we need to decide only how many rats to assign to the placebo group and how many to assign to the control group. That means we can represent our design D as $D = (n_1, n_2)$, where n_1 is the number of rats in the placebo group, and n_2 is the number in the treatment group, and $n_1 + n_2 = 20$.

In order to evaluate a design, we need to hypothesize a statistical model. Let y_{ij} be the total weight of the rat pups for the ith mother in group j. This means that y_{41} is the total weight of the pups from the fourth mother in the placebo group, and y_{42} is the total weight of the pups from the fourth mother in the treatment group. We model y_{ij} as

$$Y_{ij} = b_0 + b_j + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The quantity of interest is $b_2 - b_1$, and so we want to calculate $\text{Var}(\widehat{b}_2 - \widehat{b}_1)$ as a function of $D = (n_1, n_2)$.

Not shown here, some calculations reveal that the least squares estimator of $b_2 - b_1$ is

$$\widehat{B}_2 - \widehat{B}_1 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{i2} - \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{i1}.$$

This shouldn't be too suprising, the best estimate of the effect of the treatment is simply the average of the treatment group minus the average of the control group.

The variance of the estimator is

$$\text{Var}(\widehat{B}_2 - \widehat{B}_1) = \frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_1} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

This means that to minimize the variance of the estimator, we need to minimize $1/n_1 + 1/n_2$. We can simply enumerate all the possibilities and pick the one that minimizes the variance. A subset of the possibilities is given here:

n_1	7	8	9	10	11	12	13
n_2	13	12	11	10	9	8	7
$\frac{1}{n_1} + \frac{1}{n_2}$	0.220	0.208	0.202	0.200	0.202	0.208	0.220

We can see–also not surprisingly–that $1/n_1 + 1/n_2$ is minimized when both n_1 and n_2 are equal to 10, which tells us that the variance of the estimate of the effect of the treatment is minimized when we assign half of the female rats to the control group and half of them to the treatment group. It might be surpising that $D = (9, 11)$ and $D = (11, 9)$ are only slightly worse in terms of the variance of the estimate of the treatment effect.

Suppose had an even simpler experiment where we wanted to determine the linear relationship between the weight of the female rat and the total weight of their litter. Let y_i be the total weight of the litter for female i and x_i be the weight of female i. The simple linear model is

$$Y_i = b_0 + b_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The quantity of interest is b_1, and so we want to minimize its variance. Recall that the variance is

$$\text{Var}(\widehat{B}_1) = \frac{\sigma^2}{sxx}, \quad \text{where} \quad sxx = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Therefore, in order to give ourselves the best estimate of b_1, we should select a group of female rats with weights that are very spread out. So when you go to the animal lab, you should select female rats with a wide range of weights.

Both of these examples probably seem pretty obvious to you when you think about them, but there are good mathematical and statistical reasons supporting your intuition. In other cases, the intuition might not be so obvious, and we will need to rely on the calculations.

Chapter 11

Block Designs

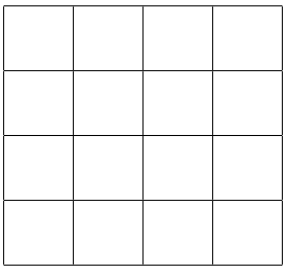
Joseph Guinness - BTRY 6020

In the last chapter, we covered some principles of experimental design, including the concepts of blocking and randomization. In this chapter, we apply these principles to outline some commonly used types of designs.

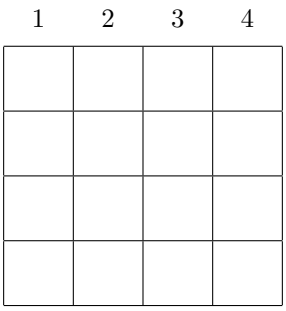
In simplest terms, a **block** is a group of experimental units. In our rat fertility example, the units are individual female rats. A block is simply a group of female rats. For example, if we divide the 20 female rats into 10 pairs of 2 rats, then each pair of two rats is a block of rats. In our fertilizer example, the units were the rows of the field. If our field has 20 rows, and we divide the 20 rows into 5 groups of 4 adjacent rows, then each group of 4 adjacent rows is technically a block. The experimenter has control over how to define the blocks.

Usually, we would like to divide our units into blocks according to some variable that we expect to influence the the response. If we do that, we can pick a design (an allocation of treatments to units) that is equal with respect to the variable that we used to define the blocks.

As an example, suppose we have four varieties of wheat (A, B, C, and D), and we would like to do a field experiment to learn about the yields of these four varieties. In this particular experiment, we can divide our field into 16 subplots (see below) and plant any variety in any subplot. This means that the 16 subplots are the units.



From our experience working with this particular field, we expect the yields to vary more in the east-west direction rather than the north-south direction (north points towards the top of the page). Knowing this information, we decide to create the blocks as follows:



so that each column in the table is a block of four subplots.

A **complete block design** (CBD) for a treatment with J levels assigns each treatment exactly once within each block. A complete block design for our wheat experiment might look like this:

1	2	3	4
A	B	B	D
B	D	A	C
C	A	C	B
D	C	D	A

Any design that has each treatment (A, B, C, D) exactly once within each block (column) qualifies as a CBD. In order to use a CBD, each block must have exactly J units. If not, then we cannot use a CBD. A **randomized complete block design** (RCBD) is simply a complete block design in which the treatments are assigned randomly to each unit within each block.

The fact that treatments appear together in the same block is important if we

want to make inferences about how the effects of the treatments differ from one another. Comparing the yield from variety A row 1 to the yield in variety C row 3 is not quite a fair comparison because we expect the yields to differ based on what row they belong to. The information in this comparison is limited by our incomplete knowledge about the effect of the row. But comparing variety A row 1 to variety C row 1 is a fair comparison. A pair of observations from the same block but different varieties contains crucial information about the effect of the variety.

The RCBD is **balanced** with respect to the treatments. We define a balanced design as one in which each treatment appears an equal number of times, and each pair of treatments appears together in the same block an equal number of times. A appears in the same block as B four times, B appears in the same block as D 4 times, and so on. We will see below that this definition of balance is not quite satisfactory, but it will suit the purpose of informing our intuitions for the time being. In balanced designs, no treatment comparison is privileged over any other treatment comparison; the study gives us the same amount of information about the difference between A and C as it does about the difference between B and D.

It is possible to have balanced designs that are not complete block designs. In some situations, the constraints of the experiment will require that the number of units within each block be less than the number of treatment levels. Suppose we could only divide our field into 12 subplots like this:

1	2	3	4

If we have four treatment levels, we cannot have a complete block design because each block contains only three units. But we can still have a balanced design:

1	2	3	4
A	D	C	B
B	A	D	C
C	B	A	D

This is an example of a **balanced incomplete block design** (BIBD). The blocks are "incomplete" because no block contains all four treatments, and the

design is balanced because each treatment appears an equal number of times (3), and each pair of treatments appears within the same block an equal number of times (2). Go ahead and verify this yourself by looking at all six pairs of treatments and counting the number of times each pair appears together within the same block.

It is not always possible to achieve a balanced incomplete block design. As an exercise, consider the scenario with four treatments, and each block contains two units. In this scenario, it is not possible to have a design wherein each pair of treatments appear together an equal number of times.

It is sometimes possible to achieve a balanced design when each block contains more units than treatment levels. For example, consider the following design where we divide our plot into four blocks of 5 units each, generating an **over-complete block design**:

1	2	3	4
A	B	B	D
B	D	A	C
C	A	C	B
D	C	D	A
A	B	C	D

Each treatment appears an equal number of times. Counting how many times each pair of treatments appears together gets a little fuzzy. Does A appear with B once or twice within block 1? This is what we meant when we said that our definition of balance is not quite satisfactory. The real definition involves ideas from abstract algebra that we can't dive into here. The basic idea of balance boils down to whether the treatments are "exchangeable" in the deisgn, in the sense that we could swap, for example, D and A everywhere in the design without changing anything meaningful about the design.

In many situations balanced designs are desirable because they minimize the average of the variance of estimators of treatment effects. But this is a good time to reiterate that achieving a balanced design is not the goal of designing an experiment. The goal is to give yourself the best chance of answering your question. Sometimes that goal is aligned with having a balanced design; sometimes it is not. For example, if you cared more about the comparisons between A and B rather than between B and C, you might place A and B together within the same block more often than you pair B with C.

As another example, suppose we had three treatments, two blocks with 3 units,

and one block with four units. We could either ignore the fourth unit in the third block, giving us a complete block design, or we could ruin the balance by planting variety A in the extra subplot:

1	2	3
A	B	C
B	C	A
C	A	B
		X

1	2	3
A	B	C
B	C	A
C	A	B
		A

For the purpose of learning about our wheat varieties, it is clearly better to have an extra observation, even if that creates an imbalance in the amount of information we have about the various treatments. It can't hurt to collect more data!

All of these examples are relatively simple because they contain only one blocking factor. In reality, there will often be more than one variable over which we want to block our treatments. In the rat example, we want to block over both the male rat and the weight of the female rat. In an agricultural field, we might expect that–regardless of treatment–the the yields will vary substantially both in the north-south and the east-west direction. The latin square design attempts to achieve balance with respect to both blocking factors.

Latin square design can be applied in the following scenario. The treatment has J levels. There are J^2 units. There are two sets of blocks, and the two sets of blocks of units are chosen so that each block from the first set of blocks contains units from all blocks of the second set of blocks (and vice versa). The latin square design places exactly one treatment of each level in each block. This is most easily explained in diagrams. Suppose that our field has 16 subplots, and we define rows to be one set of blocks, and columns to be the other set of blocks.

Not a Latin square

	1	2	3	4
1	A	B	B	D
2	B	D	A	C
3	C	A	C	B
4	D	C	D	A

Latin square

	1	2	3	4
1	A	B	D	C
2	B	C	A	D
3	C	D	B	A
4	D	A	C	B

The two sets of blocks (rows and columns) satisfy our requirement that each block of the first factor contains units from every block of the second factor. More concretely, each row contains a unit from each column, and each column contains a unit from each row. The design on the left is not a latin square because, even though each treatment appears exactly once within each column, B appears twice in the first row, and C appears twice in the third row. The design on the right is a latin square because each treatment appears exactly once within each row (the first set of blocks) and exactly once within each column (the second set of blocks).

Chapter 12

Factorial Designs

Joseph Guinness - BTRY 6020

The block designs in the previous chapter are feasible when the total number of units is the same or larger than the number of possible treatment combinations. When there are several treatment factors with several levels each, the number of possible treatment combinations can be very large. For example, suppose we want to study how a projection from a computationally expensive climate model depends on 6 different input parameters, each of which can be set to four different levels. Our budget allows us to run the model 16 times, yet there are $4^{10} = 4096$ possible combinations of these 10 factors.

Factorial designs, and especially fractional factorial designs, are solutions to the problem of having many possible combinations of treatments. In a **factorial design**, each treatment factor is limited to two levels, a "low" and a "high" setting, and then the experiment is run under all possible settings. In a **fractional factorial design**, the experiment is run under a specific subset of all possible settings.

Suppose there are k treatment variables. For each treatment variable, we identify a low and a high value, and we view each of the k treatments as a factor with $J = 2$ levels. This means there are a total of 2^k possible settings. For example, if we have $k = 2$ treatment variables, and we label the low settings as A and the high settings as B, these are the four possible settings for the experiment:

Setting	Fac 1	Fac 2
1	A	A
2	A	B
3	B	A
4	B	B

Note that factor 1 at its low setting and factor 2 at its high setting is different from factor 1 at its high setting and factor 2 at its low setting. Hence rows 2 and 3 are separate possible settings.

For $k = 3$, there are $2^3 = 8$ possible settings:

Setting	Fac 1	Fac 2	Fac 3
1	A	A	A
2	A	B	A
3	B	A	A
4	B	B	A
5	A	A	B
6	A	B	B
7	B	A	B
8	B	B	B

When $k = 4$, there are 16 possible settings; we won't list them all here.

The 2^k **factorial design** or **full factorial design** in k two-level factors is simply an experiment that is run at all 2^k combinations of the factor levels. So a 2^2 factorial design needs 4 units, a 2^3 factorial design needs 8 units, and a 2^4 factorial design needs 16 units.

As you can see, the number of units required for the 2^k factorial design grows quickly (exponentially even!) with k. If we have 6 treatment factors, we need $2^6 = 64$ units for a full factorial design. This is simply not feasible for our climate model study. However, we still want to investigate the effect of each of the k variables. The solution is to run the experiment at a subset of all of the 2^k possible combinations of the treatments. This is called a **fractional factorial design**.

The **one-half fraction factorial design** of the 2^k factorial design contains half of the runs of a 2^k factorial design. These $2^k/2 = 2^{k-1}$ runs are chosen so that if any varible is removed, we obtain a $2^{(k-1)}$ design in the remaining $k - 1$ variables. The following table explains the notation for referring to these designs

2^{k-1}	Full factorial design in $k - 1$ variables
$2^{(k-1)}$	One half fractional factorial design in k variables

This is a big tough to swallow, so let's do some examples. Here are two possible designs in 2 factors that have 2 runs each:

Design 1

Run	Fac 1	Fac 2
1	A	B
2	B	A

Design 2

Run	Fac 1	Fac 2
1	A	B
2	A	A

One of these is a 2^{2-1} design, and the other is not. Use your finger to cover either the factor 1 column or the factor 2 column. If the design includes both levels of the other factor, it is a 2^{k-1} design. Let's see how this works. In Design 1, if you cover factor 1, you have both settings of factor 2, and if you cover factor 2, you have both settings of factor 1. Therefore, Design 1 is a 2^{2-1} design. In Design 2, if you cover factor 1 you have both settings of factor 2, but if you cover factor 2, you have only one setting of factor 1. Therefore, Design 2 is not a 2^{2-1} design.

The following is also a valid 2^{2-1} design:

Run	Fac 1	Fac 2
1	A	A
2	B	B

You can verify that for youself by covering either column with your finger. These are the only two 2^{2-1} designs. In fact, there are only two 2^{k-1} designs for each k.

Let's do the $k = 3$ example. These are the two 2^{3-1} designs

Run	Fac 1	Fac 2	Fac 3
1	A	B	A
2	B	A	A
3	A	A	B
4	B	B	B

Run	Fac 1	Fac 2	Fac 3
1	B	A	B
2	A	B	B
3	B	B	A
4	A	A	A

If you place your finger over any of the three factor columns, the design contains all possible combinations of the remaining two factors.

Fractional factorial designs are particularly useful in situations where you have many treatments, but you expect that some of them do not influence the response, but you don't know which ones. You can run the experiment, figure out which ones didn't influence the response by fitting an additive model in each factor, then remove the unimportant variables and fit more complicated interaction models in the remaining variables. When you remove variables, you have to be careful not to overinterpret the significance tests, though, because the model has been chosen based on peeking at the data.

We can take the idea of fractional factorial designs further. There exist 2^{k-2}

fractional factorial designs, which are one-quarter fractions of the 2^k design. These designs are constructed so that if any two variables are removed, the design is a full $2^{(k-2)}$ design in the remaining $k-2$ variables. This is an example of a 2^{5-2} design:

Run	Fac 1	Fac 2	Fac 3	Fac 4	Fac 5
1	A	A	A	B	B
2	B	A	A	A	A
3	A	B	A	A	B
4	B	B	A	B	A
5	A	A	B	B	A
6	B	A	B	A	B
7	A	B	B	A	A
8	B	B	B	B	B

If you cover any two columns, the design is a full factorial design in the remaining three columns. This allows us to explore the effects of 5 factors with only 8 rather than $2^5 = 32$ runs of the experiment.

Since we have 6 factors in our climate model study, but can do only 16 runs, we could do a 2^{6-2} fractional factorial design:

Run	Fac 1	Fac 2	Fac 3	Fac 4	Fac 5	Fac 6
1	A	A	A	A	A	A
2	B	A	A	A	B	A
3	A	B	A	A	B	B
4	B	B	A	A	A	B
5	A	A	B	A	B	B
6	B	A	B	A	A	B
7	A	B	B	A	A	A
8	B	B	B	A	B	A
9	A	A	A	B	A	B
10	B	A	A	B	B	B
11	A	B	A	B	B	A
12	B	B	A	B	A	A
13	A	A	B	B	B	A
14	B	A	B	B	A	A
15	A	B	B	B	A	B
16	B	B	B	B	B	B

Chapter 13

Random Effects Models

Joseph Guinness - BTRY 6020

Before introducing our first random effects model, let's take a moment to review the one factor model and all of its assumptions. Suppose we have a single factor with J levels, and we write y_{ij} as the ith response that had factor level j. We modeled y_{ij} as

$$Y_{ij} = b_0 + b_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

On the left side of the equation we have a random variable Y_{ij}. On the right side of the equation there are three terms: b_0, b_j, and ε_{ij}. Other than assuming that b_0 and b_j are numbers, there are no assumptions placed on them. They could be anything, and further, no assumptions are made about the *collection* of numbers $b_1, \ldots, b_J$. They could be any collection of numbers. We do place an assumption on ε_{ij}, namely that it follows a normal distribution with mean zero and variance σ^2. The i.i.d. assumption is a further assumption about the collection of variables $\varepsilon_{11}, \varepsilon_{21}, \ldots, \varepsilon_{12}, \ldots$, namely that they are independent and all follow the same normal distribution.

The one-factor model above is sometimes called a **fixed effects model**, owing to the fact that the effect of changing the factor level, codified with the parameters $b_1, \ldots, b_J$, is a nonrandom, or fixed, quantity.

As opposed to fixed effects models–where we place no assumptions on the collection of effects $b_1, \ldots, b_J$–in **random effects models**, we do make an assumption about the collection of effects. There are various types of random effects models, but the most commonly used ones, and the ones we will study in this chapter, place a normal assumption on the effects. In particular, the random effects version of the one factor model is

$$Y_{ij} = b_0 + B_j + \varepsilon_{ij}, \quad B_j \overset{iid}{\sim} N(0, \sigma_1^2), \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_2^2).$$

As before, the left side of the equation is a random variable Y_{ij}. On the right side of the equation are three terms: b_0, B_j, and ε_{ij}. No assumptions are placed on the first term, b_0. The second term B_j, is modeled with a normal distribution with mean zero and variance σ_1^2. Moreover, the collection of random variables $B_1, \ldots, B_J$ are independent and all follow the same normal distribution. The same assumptions as before are placed on the collection of variables ε_{ij}.

The notation for the random effects model capitalizes B_j to help remind us that now B_j is a random variable. The most important thing to remember about random effects models is that they place an assumption on the collection of effects $B_1, \ldots, B_J$. In this particular model, we assume they are normally distributed, meaning that if our factor had many levels, i.e. J were large, and we were able to make a histogram of $B_1, \ldots, B_J$, then the histogram would follow a bell curve. On the other hand, no such assumption is made about $b_1, \ldots, b_J$ in the fixed effects models. In that model, we do not make any prior judgments about the distribution of the effects. Their histogram could take on any shape.

There are many types of random effects models. We should view all of them as making stronger assumptions about the model than their fixed effect counterparts.

Since the random effects model places an assumption on a collection of variables, it makes the most sense to use random effects models in the context of factor variables that have multiple levels. If we have two factors with responses y_{ijk}, belonging to the ith response from level j of factor 1 and level k of factor 2, we could make one of the factors' effects random, as in

$$Y_{ijk} = b_0 + B_j + c_k + \varepsilon_{ijk}, \quad B_j \overset{iid}{\sim} N(0, \sigma_1^2), \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2).$$

Or we could make both of factors' effects random, as in

$$\begin{aligned} &Y_{ijk} = b_0 + B_j + C_k + \varepsilon_{ijk}, \\ &B_j \overset{iid}{\sim} N(0, \sigma_1^2), \quad C_k \overset{iid}{\sim} N(0, \sigma_2^2), \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2). \end{aligned}$$

If we have a factor with J levels and a numeric covariate, we could write down an additive factor-numeric model in which the intercepts are random, as in

$$Y_i = a_0 + A_{j(i)} + b_0 x_i + \varepsilon_i, \quad A_j \overset{iid}{\sim} N(0, \sigma_1^2), \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_2^2).$$

In this model, the collection of intercepts $a_0 + A_1, \ldots, a_0 + A_J$ follows a normal distribution with mean a_0 and variance σ_1^2. If we want interactions between the factor and numeric covariate, we can model the level-dependent slopes as random also:

$$\begin{aligned} &Y_i = a_0 + A_{j(i)} + (b_0 + B_{j(i)}) x_i + \varepsilon_i, \\ &A_j \overset{iid}{\sim} N(0, \sigma_1^2), \quad B_j \overset{iid}{\sim} N(0, \sigma_2^2), \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_3^2). \end{aligned}$$

Here, the collection of intercepts follows a normal distribution, and the collection of slopes follows a different normal distribution.

13.1 When should we use a random effects model?

In answering this question, let me reiterate that when we make some of the parameters random, we are placing an additional assumption on the model. The thought process for choosing a random effects model versus a fixed effects model is the same thought process used for any other assumption you make in a statistical analysis:

1. Is the assumption is close to correct?
2. Do the potential benefits of making the assumption outweigh the potential costs of the assumption when it's wrong?

You might be wondering why we would ever make an assumption when we don't have to. This is a reasonable question to ask. The answer is that there are benefits to making assumptions. When I drive to work, I usually assume that the roads I have taken in the past still exist and can be traveled without major impediment. This assumption buys me some saved time in the morning, in that I don't have to seek out information about the condition of the roads. If I lived in a more congested area, this assumption is more likely to be wrong, and could cost me. Living in a congested area, I would probably have to watch the local news in the morning to check the traffic conditions.

Making a (near) correct assumption about terms in a random effects model can benefit us by giving us more precise estimates of quantities we care about. For example, if we correctly assume that male rat effects are normally distributed, we can get more precise estimates of the effect of a hormone. Random effects models also allow us to make predictions. For example, if we make no assumptions about the collection of male rat effects, we have no basis for predicting the effect of a new male rat. On the other hand, if we assume that male rat effects follow a normal distribution with variance σ_1^2, we can get an estimate of σ_1^2 from the data, and use the estimated normal distribution to predict the effect of the next male rat.

13.2 Estimating Random Effects Models

Random effects models are the first situation we have encountered where we don't usually use the least squares criterion to estimate parameters in the model. The details for how parameters are estimated in random effects models are beyond the scope of this course. We usually use a technique called maximum likelihood estimation or residual maximum likelihood estimation, also known as restricted maximum likelihood estimation. It suffices to say that the least squares estimates of the parameters in the fixed effects model are based on the likelihood function of the fixed effects model, and likewise the estimates of the parameters in random effects model are based on the likelihood function for the random effects model. So there is a deeper connection between the

estimation techniques, but the model is different when the effects are assumed to be random.

13.3 Modeling Dependence Using Random Effects

In all of the fixed effects models that we've studied this semester, we have assumed that every observation is independent of every other observation. Random effects models offer us a way to build dependence into the model when it is appropriate.

Consider a fixed effect model for y_{ij}, the score for student i on quiz j, in which quiz number is a factor covariate:

$$Y_{ij} = b_0 + c_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

We can also write this model as

$$Y_{ij} \overset{ind}{\sim} N(b_0 + c_j, \sigma^2).$$

The one factor model is probably not adequate for the quiz scores because we expect similarity between scores from different quizzes that come from the same student. We now know of two models to remedy the situation:

$$\begin{aligned} Y_{ij} &= b_0 + b_i + c_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \\ Y_{ij} &= b_0 + B_i + c_j + \varepsilon_{ij}, \quad B_i \overset{iid}{\sim} N(0, \sigma_1^2), \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_2^2). \end{aligned}$$

In the first model–the fixed effects model–each student i has its own nonrandom effect b_i. The same is true for the random effects model, except that the collection of student effects $B_1, B_2, \ldots$ is assumed to follow a normal distribution.

Let us further analyze the random effects model. The response Y_{ij} is the sum of a random term, a normal random variable, and another normal random variable, independent of the first. Therefore Y_{ij} has a normal distribution. Its expected value and variance are

$$\begin{aligned} E(Y_{ij}) &= E(b_0 + B_i + c_j + \varepsilon_{ij}) = b_0 + E(B_i) + c_j + E(\varepsilon_{ij}) = b_0 + c_j \\ \mathrm{Var}(Y_{ij}) &= \mathrm{Var}(b_0 + B_i + c_j + \varepsilon_{ij}) = \mathrm{Var}(B_i) + \mathrm{Var}(\varepsilon_{ij}) = \sigma_1^2 + \sigma_2^2, \end{aligned}$$

and therefore

$$Y_{ij} \sim N(b_0 + c_j, \sigma_1^2 + \sigma_2^2),$$

which might seem very similar to the quiz-only model. The difference here is that in the random student effect model, quiz scores from the same student are

assumed to be correlated. Recall that correlation between random variables A and B is defined as

$$\text{Corr}(A, B) = \frac{\text{Cov}(A, B)}{\sqrt{\text{Var}(A)\text{Var}(B)}},$$

and covariance is defined as

$$\text{Cov}(A, B) = E\Big(\big[A - E(A)\big]\big[B - E(B)\big]\big)\Big).$$

We can use these formulas to determine the correlation between quiz scores. For this example, let's assume we are looking at student 6's scores on quizzes 1 and 2. Going through the calculation line by line gives

$$\begin{aligned}
\text{Cov}(Y_{61}, Y_{62}) &= E\Big(\big[b_0 + B_6 + c_1 + \varepsilon_{61} - b_0 - c_1\big]\big[b_0 + B_6 + c_2 + \varepsilon_{62} - b_0 - c_2\big]\Big) \\
&= E\Big(\big[B_6 + \varepsilon_{61}\big]\big[B_6 + \varepsilon_{62}\big]\Big) \\
&= E\Big(B_6B_6 + B_6\varepsilon_{62} + \varepsilon_{61}B_6 + \varepsilon_{61}\varepsilon_{62}\Big) \\
&= E(B_6^2) + E(B_6\varepsilon_{62}) + E(\varepsilon_{61}B_6) + E(\varepsilon_{61}\varepsilon_{62}) \\
&= \text{Var}(B_6) + E(B_6)E(\varepsilon_{62}) + E(\varepsilon_{61})E(B_6) + E(\varepsilon_{61})E(\varepsilon_{62}) \\
&= \sigma_1^2 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 \\
&= \sigma_1^2.
\end{aligned}$$

Before we move on, note that in these calculations $E(B_6\varepsilon_{61}) = E(B_6)E(\varepsilon_{61})$ because B_6 and ε_{61} are independent. Likewise for B_6 and ε_{62} and for ε_{61} and ε_{62}. However, $E(B_6B_6) \neq E(B_6)E(B_6)$ because B_6 is very clearly not independent of itself. Converting $E(B_6^2)$ to $\text{Var}(B_6)$ uses the definition of variance:

$$\text{Var}(A) = E\Big(\big[A - E(A)\big]^2\Big) \quad \Longrightarrow \quad \text{Var}(B_6) = E([B_6 - 0]^2) = E(B_6^2).$$

From the calculation above, we learned that the covariance between Y_{61} and Y_{62} is σ_1^2. The variance of each variable is $\sigma_1^2 + \sigma_2^2$. Putting this all together, the correlation between Y_{61} and Y_{62} is

$$\text{Corr}(Y_{61}, Y_{62}) = \frac{\text{Cov}(Y_{61}, Y_{62})}{\sqrt{\text{Var}(Y_{61})\text{Var}(Y_{62})}} = \frac{\sigma_1^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(\sigma_1^2 + \sigma_2^2)}} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}.$$

This means that the larger the variance of B_6 relative to the variance of ε_{ij}, the larger the correlation.

It took some laborious calculations to get to this point, but to summarize, quiz scores from the same student are assumed to have positive correlation. On the other hand, quiz scores from different students have zero covariance,

$$\text{Cov}(Y_{61}, Y_{72}) = 0,$$

and therefore zero correlation.

13.3.1 Multilevel Models

Suppose that an education researcher is studying a new method for math education in middle schools. She recruits a number of teachers for the study. The teachers come from a number of different schools, with some teachers coming from the same school as other teachers. Each teacher is assigned to either a control group or the treatment group. Teachers in the treatment group attend a day-long workshop being trained in the new methods. The teachers in the control group also attend a day-long training workshop, but instead freshen up on traditional math teaching methods (Think about how you would assign teachers to treatment vs. control knowing that some teachers come from the same school.)

In order to collect data for the project, each teacher gives his or her students a pre-test at the beginning of the study. Teachers in the treatment group deliver instruction in the new methods for 6 weeks, while teachers in the control group carry on as usual. At the end of 6 weeks, the students are given a post-test. The data we analyze are y_{ijk} the difference the post-test and pre-test score for student k in teacher j's classroom from school i. Note that teacher 2 in school 3 is different from teacher 2 in school 4.

We will need to descend into subscript hell in order to write down a mathematically coherent model for these data. To set the stage, define

$$\begin{aligned} t(i,j) &= 1 \quad \text{if teacher } j \text{ from school } i \text{ got the control,} \\ t(i,j) &= 2 \quad \text{if teacher } j \text{ from school } i \text{ got the treatment.} \end{aligned}$$

We model the responses as

$$\begin{aligned} Y_{ijk} &= b_0 + b_{t(i,j)} + C_i + D_{ij} + \varepsilon_{ijk}, \\ C_i &\overset{iid}{\sim} N(0, \sigma_1^2) \\ D_{ij} &\overset{iid}{\sim} N(0, \sigma_2^2) \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma_3^2). \end{aligned}$$

Note that we use D_{ij} instead of D_j because teacher j from school 1 is different from teacher j in school 2, so those two teachers should not share a random effect (this is a very common mistake in notating multilevel models).

Even though we have a complicated notation for the fixed effects $b_{t(i,j)}$ there are really only two of them, b_1 and b_2 since $t(i,j)$ can only be 1 or 2. The goal of the study is to make inferences about the efficacy of the new teaching method. In our model, that efficacy is encoded as $b_2 - b_1$, the differences in expected improvement in the treatment group versus the control group.

The other terms in the model, C_i and D_{ij}, capture the fact that improvements for two students that share a teacher should be correlated, due to the strength or weakness of their common teacher. Likewise, improvements from two students

in the same school should be correlated, due to the strength of the school, or common experiences shared by students from the same school. We should also expect that students that share a teacher are more correlated than students who share a school but not a teacher.

Let's do some calculations to see how the model realizes these intuitive assumptions about correlation. The variance of each observation is $\sigma_1^2 + \sigma_2^2 + \sigma_3^2$.

Two students who share a teacher:

$$\begin{aligned}\mathrm{Cov}(Y_{731}, Y_{732}) &= E\Big([C_7 + D_{73} + \varepsilon_{731}][C_7 + D_{73} + \varepsilon_{732}]\Big) = \sigma_1^2 + \sigma_2^2 \\ \mathrm{Corr}(Y_{731}, Y_{732}) &= \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}.\end{aligned}$$

Two students who share a school but not a teacher:

$$\begin{aligned}\mathrm{Cov}(Y_{731}, Y_{742}) &= E\Big([C_7 + D_{73} + \varepsilon_{731}][C_7 + D_{74} + \varepsilon_{742}]\Big) = \sigma_1^2 \\ \mathrm{Corr}(Y_{731}, Y_{742}) &= \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}.\end{aligned}$$

Two students in different schools:

$$\begin{aligned}\mathrm{Cov}(Y_{731}, Y_{542}) &= E\Big([C_7 + D_{73} + \varepsilon_{731}][C_5 + D_{54} + \varepsilon_{542}]\Big) = 0 \\ \mathrm{Corr}(Y_{731}, Y_{542}) &= \frac{0}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2} = 0.\end{aligned}$$

In some situations we might be directly interested in knowing these components of the variance σ_1^2, σ_2^2, and σ_3^2, or in knowing the correlations. But even if we are only interested in knowing the treatment effect $b_2 - b_1$, we still need to account for the school and teacher effects because we expect students that share a teacher or share a school to have **dependent** improvements. Random effects models are a way to incorporate dependence into the model.

13.4 Direct Modeling of Dependence

There are many applications where the random effects models studied so far are suitable for capturing dependence between observations. However, there are other situations where these types of random effects models are inadequate, and we need a different strategy for modeling dependence. The strategy we have employed so far is to write down a model equation with random terms that are shared by observations that we aim to model as dependent. We had a random term for student in the quiz model because we expected quiz scores from the same student to be dependent. We have a random term for school in the education experiment because we expected students from the same school

to be dependent. In these models, the covariance between observations came as a consequence of these shared random effects terms.

A different strategy altogether is to directly specify the covariances between observations. As a prime example, consider an agricultural trial in which we expect the fertility of the soil to be a smooth function of spatial locations. If this is the case, then we expect yields from subplots i and j to be similar if subplots i and j are close together. So it seems reasonable to model the yields as

$$\begin{aligned} Y_i &\sim N(b_0 + b_{t(i)}, \sigma^2) \\ \text{Cov}(Y_i, Y_j) &= f(d_{ij}) \\ f(d_{ij}) &= \sigma^2 e^{-d_{ij}/a}, \end{aligned}$$

where $t(i)$ is the treatment assignment, and d_{ij} is the distance between subplots i and j. In this model, the covariance between yields decays exponentially with the distance between the subplots.

This strategy is more general than the previous one because we wouldn't be able to generate a model like this by assigning random effects to different parts of the field. We do have to be careful about how we specify the covariances. The covariance function $f(d_{ij})$ has to satisfy a property called **positive definiteness**, so not every function you dream up will work. The exponential happens to satisfy this property, and there are a host of others if you think that the exponential is not appropriate for your data.

Chapter 14

Generalized Linear Models

Joseph Guinness - BTRY 6020

All of the models that we have studied so far assume that the responses follow a normal distribution. There are multiple ways of writing these models, for example, if the expected value is a linear function of two covariates, we can write the model for the response y_i as

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Equivalently, we could define the model for y_i in terms of its distribution, as

1. $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$
2. $\mu_i = b_0 + b_1 x_{i1} + b_2 x_{i2}$.

Here, we've split the specification of the linear mean function into a separate part, but's all still the same.

We've carefully studied different specifications of the mean part of the model μ_i– polynomials, factors, additive models, and interactions. We have also studied ways to induce dependence between groups of observations via random effects. The normal model has taken us a long way, and there are many important datasets for which this type of model is reasonable.

However, the normal model is clearly inappropriate for some types of data. As a prime example, suppose we are studying a new treatment for improving resistance to invasive tree beetles, and the response is either 1 if the beetles infect the tree, and 0 if not. A binary variable is not normally distributed, and moreover, there is no transformation of 0 and 1 that would make them normal–the resonse will always be discrete with two possible values, far from a normal distribution. For this type of data and others, we need a new modeling framework.

The generalized linear model is one way of modeling data with a variety of distributions. The generalized linear model has 3 parts:

$$\begin{aligned} \text{Family:} \quad & Y_i \overset{ind}{\sim} \text{Distribution}(\mu_i) \\ \text{Link:} \quad & \mu_i = g(a_i) \\ \text{Linear Part:} \quad & a_i = b_0 + \sum_{j=1}^{p} b_j x_{ij}. \end{aligned}$$

You'll notice that the specification of the generalized linear model resembles our specification of the normal model above, but with some key differences. First, the distribution is potentially not normal. In a specific generalized linear model, we will replace "Distribution" by "Binomial", "Poisson", "Gamma", etc. Second, the parameter in the distribution is not a linear function of the covariates; instead the parameter is a function of a_i, and a_i is a linear function of the covariates $x_{i1}, \ldots, x_{ip}$. This *link function* g will depend on what the modeler chooses and will vary from problem to problem. We will see soon why the link function is necessary. The main reason is that for non-normal distributions, the parameter is sometimes constrained to fall within a specific interval. In the binomial, the prameter is a probability and therefore must lie between 0 and 1. In the Poisson, the parameter must be a positive number.

Additionally, in the specification of the linear part of the model, there is no random error term. The "noise" in the data comes from the fact that the data are assumed to be independent. Do not make the mistake of writing a generalized linear model with an additive error term (unless you are working with generalized linear mixed models).

Let's look at some specific examples. To keep things simple, we will assume that there is just a single covariate. In real examples, we often have several covariates, just like in the normal models we have studied.

Bernoulli Model (Logistic Regression)

The Bernoulli random variable can take on the values 0 or 1. To specify the distribution, we simply need to say that $P(Y_i = 1) = p_i$, and therefore $P(Y_i = 0) = 1 - p_i$. So the parameter of interest is p_i. The generalized linear model for the Bernoulli distribution is

$$\begin{aligned} \text{Family:} \quad & Y_i \overset{ind}{\sim} \text{Bernoulli}(p_i) \\ \text{Link:} \quad & p_i = \frac{e^{a_i}}{1 + e^{a_i}} \\ \text{Linear Part:} \quad & a_i = b_0 + b_1 x_i. \end{aligned}$$

It is common to use the link function $\exp(a_i)/(1 + \exp(a_i))$ because no matter what a_i is, p_i must be between 0 and 1, which is desirable because p_i is a

probability. To see why, note that $\exp(a_i)$ must be a positive number less than $1 + \exp(a_i)$.

This model is sometimes called the logistic regression model. This comes from the fact that a_i can be expressed in terms of p_i as

$$a_i = \log\Big(\frac{p_i}{1-p_i}\Big),$$

which is known as the *logistic function* of p_i.

Binomial Model

The binomial model is a model for counts that must lie between 0 and some known upper limit. For example, in our tree beetle study, suppose instead that each "observation" is the total number of infected trees within each study area, and each study area has a known number of total trees. The binomial model is specified in terms of a probability p_i (representing the probability that each individual tree is infected), and m_i, the total number of trees, and has probability mass function

$$P(Y_i = k) = \frac{m_i!}{k!(m_i-k)!} p_i^k (1-p_i)^{m_i-k}, \quad k = 0, 1, 2, \ldots, m_i.$$

An example of a binomial gneneralized linear model is

$$\begin{aligned} \text{Family:} \quad & Y_i \overset{ind}{\sim} \text{Binomial}(p_i, m_i) \\ \text{Link:} \quad & p_i = \frac{e^{a_i}}{1+e^{a_i}} \\ \text{Linear Part:} \quad & a_i = b_0 + b_1 x_i. \end{aligned}$$

This is very similar to the Bernoulli model. In fact, the binomial model is a Bernoulli model when $m_i = 1$.

Poisson Model

The Poisson model is a model for unbounded counts. For example, suppose we wanted to model the number of shooting stars observed on every night for a year. The total number of shooting stars could essentially be any natural number. The Poisson model is also sometimes used when the observed count is expected to be much less than the theoretical maximum count. For example, supposing beetle infestations are rare, we might model the number of infested trees in a large tract of land as Poisson, even though the number infested cannot exceed the total number of trees.

The poisson random variable has a positive parameter μ_i representing the expected value, $E(Y_i) = \mu_i$. The probability mass function is

$$P(Y_i = k) = \frac{e^{-\mu_i}\mu_i^k}{k!}, \quad k = 0, 1, 2, 3, ...$$

An example of a generalized linear model is

$$\begin{aligned} \text{Family:} \quad & Y_i \overset{ind}{\sim} \text{Poisson}(\mu_i) \\ \text{Link:} \quad & \mu_i = \exp(a_i) \\ \text{Linear Part:} \quad & a_i = b_0 + b_1 x_i. \end{aligned}$$

Since the Poisson is a model for counts (non-negative numbers) the expected value must be a positive number, and therefore $\exp(a_i)$ is a reasonable link function, though others are possible.

Exponential Model

The previous models are all models for counts, which is the most obvious application for a generalized linear model, since the normal distribution is continuous, whereas counts are not. Nevertheless, there are situations where the response is continuous (could take on any real number) but the normal is not an appropriate model. A good example is in survival analysis, when we aim to model the length of time a subject survives, as a function of their disease status or treatment.

One popular response distribution is the exponential distribution, which has probability density function

$$f(y) = \begin{cases} \frac{1}{\mu}\exp(-y/\mu) & y \geq 0 \\ 0 & y < 0. \end{cases}$$

The parameter μ represents the expected survival time, $E(Y) = \mu$. A generalized linear model is

$$\begin{aligned} \text{Family:} \quad & Y_i \overset{ind}{\sim} \text{Exponential}(\mu_i) \\ \text{Link:} \quad & \mu_i = \exp(a_i) \\ \text{Linear Part:} \quad & a_i = b_0 + b_1 x_i. \end{aligned}$$

Again, we use the exponential link, though others are acceptable. Depending on the application, we might select covariates that allow us to model the log of the expected survival time as a linear function of treatment, age, or disease severity.

There are several other responses distributions that are commonly used, but we cannot cover them all in detail. The negative binomial is another count distribution, and the gamma is another continuous distribution.

The parameters in the generalized linear model consist of $b_0, b_1, \ldots, b_p$, which describe the linear relationship between the covariates and a_i. These parameters are estimated via the maximum likelihood method, the details of which are not too difficult to understand but are a bit too complicated to cover in this course. In R, we use the `glm` function to fit the models.

Bibliography

[1] Peter McCullagh. What is a statistical model? Annals of statistics, pages 1225–1267, 2002.