QERM 514, Class Notes

Spring, 2008

Hans Nesse

Version Final, compiled May 31, 2008

An addax, *Addax nasomaculatus*, from Ernst Haeckel's *Kunstformen der Natur* (1904), currently classified as endangered on the IUCN redlist.

Contents

-1.1 Syllabus

This is a course giving an overview of some of the statistical tools which frequently crop up in an ecological setting. Statistics is an entire discipline, of course, and there are a great many specializations. While no scientist can afford to ignore statistics entirely, being a master of all statistical methods which might come up is an unrealistic goal for any scientist (or really even a statistician). Likewise the goals of this course can not be to go over all methods which are likely to be relevant–there is simply too much statistical information developed and commonly applied.

For that reason, this course is limited to a subset of the range of useful statistics, but hopefully a careful and complete enough introduction to these methods that extension into other fields becomes easier.

-1.1.1 Dramatis personae

There are three people responsible for QERM 514 this year. The faculty member who is responsible for the class is Dr. Vince Gallucci, however he will not be giving lectures or working lab sessions[1]. Those responsibilities fall to two QERM graduate students, Hans Nesse and Derek McClure. Hans is delivering the lectures and grading homework, while Derek runs optional lab sessions. Questions and concerns about course content should be primarily directed to either Hans or Derek first.

Hans Nesse	Derek McClure
Office Ph: 206.221.6776	206.685.4492
Office: FSH 262A	Lab: FSH 209
Hours: W 10-12pm	Lab: Tu 12.30-2.30
Or by appt.	
Email: `nesse@ u.`	`derekm3@u.`

This course was developed out of (that is, copied from) a course taught by Dr. Loveday Conquest.

-1.1.2 Notes and other resources

These notes comprise the principle source of content for this course. Lectures are intended to follow chapter outlines fairly closely. However the exposition in these notes is fairly limited. As such, two additional books are recommended texts: *Linear Models with* `R` and *Extending the Linear Model with* `R`, both by Julian Faraway, are recommended for a clear exposition of the topics in this course (and many others besides). There are also a list of texts in section -1.3 which may be useful for some topics, most of which are available from the UW library. Don't hesitate to ask for additional resources if you feel it could be beneficial.

In a sense, these notes are written backwards–the examples are entirely at the end of the lecture. In statistics, however, it is often useful to start with concrete examples before moving to developing theory. The ordering is intended to follow the lecture, which is largely devoted to theory, however these notes may be difficult to read as a textbook. Sometimes the notes are more of an appendix, treating in detail what is briefly covered in class.

[1] For QERM students, however, he is responsible for writing part of the QERM Applied Qualifying exam.

The notes should be considered a draft at any point until they are presented. Many are missing examples, and may need corrections. A copy of the most current copy of the days notes will be distributed in class the day of the lecture.

A special note about proofs

This course does not prove everything presented, however there are many proofs in (or at least sketched) the notes. Going over a proof is a good way to reinforce a concept and ensure that all the details are clear; however some people in this course may have difficulty with them. The balance that is struck is to present proofs but not require their use or comprehension for assignments or exams (with a few exceptions). Note, however, that this only means that the assignments won't directly rely on understanding the proof, not that there is no benefit to learning them.

Matrices and statistical methods

This course unabashedly uses both matrix notation (when convenient) and appeals to some level of understanding in probability. The course is intended to be an introduction to further reading in applied statistics, using the language of those works is important. Matrix notation (and especially multivariable calculus in matrix notation) is often a point of difficulty. In a sense, it is a topic in its own right in the course (although one which receives little specific attention in lecture). In the long run, this is perhaps as important as any other topic in the course, since it furthers a reading knowledge of additional methods. But it can be initially frustrating. Much of the use of matrices is in the first part of the course, and there are many sources which cover these topics without appealing to matrix notation.

-1.1.3 Prerequisites

There are no official prerequisites for the class, however it is a graduate-level course and is a core part of the QERM first year. As such, it may be inaccessible to some students with limited backgrounds in either statistics or mathematical methods. It is not intended to be a first course in statistics, however it is not (entirely) theoretically oriented either. A reasonable background needed to take the class would be minimally, a course in statistics, elementary multivariable calculus (with a few exceptions in proofs, freshman calculus should be fine), and some facility with matrix algebra. Statistical and probability background should certainly include random variables, distributions, hypothesis testing (ideas like p-values, null and alternative hypotheses, etc.), and some idea of point estimation.

-1.1.4 Grading

There are three components to the grade in the course: Homework, exams, and presentations. All of these are on the same point scale (point values for each are given below). Chances are grades are not the motivating factor of taking this class, but they must be assigned.

Homework

Homework is assigned weekly and are each worth 10 points. Homework is usually in the form of a statistical analysis. In some cases, more than one method may be appropriate. The intent is to

mirror, on a small scale, the statistical portion of the QERM applied qualifying exam. Although there is theoretical content covered in class and in the notes, it is the application that is foremost. All homework should be written as a report: results should not simply be given, but also described. Please also include an appendix for your code.

Exams

The exams, one midterm and a final, are intended to be slightly larger versions of the homework. Both are take-home and open book (any book) and notes, and are worth 25 points each. While collaboration and discussion is encouraged on the homework, please do the exams without getting help from other people (other than Hans or Derek).

Presentation

Early in the class, a sign-up sheet will be circulated for topics. Everyone should sign up to present examples for two topics, which is 15 points each. One you have selected topics, you will need to (i) find an example for each topic and (ii) work the example, (iii) present it in class, and (iv) turn in a hand-out (or make copies for everyone) detailing and summarizing the results of your analysis.

A student signed up for a topic is expected to do some independent reading on it and develop an example problem illustrating the topic, and present it in class. A brief write-up of the problem should also be turned in, however a broader exposition of the topic is strongly encouraged.

Grades on presentations, like homework, are based on clarity of presentation (both in class and written), and work required (particularly if you use real data). You should start researching this early, and come see either Derek or Hans if you have trouble understanding a topic or finding an example. *Unless you specifically request otherwise, a copy of your write-up will be distributed to other students for their reference, after your presentation.* (You don't need to put together a powerpoint presentation, but if you do and need a projector, please be sure to bring one; the room is not equipped for such high-tech-ness.)

A 8-12 minute presentation is the ideal length, although an interesting problem trumps time anyday. If possible, turn in the written part of the presentation at least 24 hours ahead of time, but certainly no later than the day of the presentation. If you're unsure of your analysis, feel free to discuss it with Hans or Derek before the day you are to present.[2]

Extra credit

Extra credit is available for working specific problems, and will be handled in an unusual manner. Some extra credit assignments are available on the course website. However, if you come up with an interesting question which is topically related to the class, you may do it for extra credit of up to 5 points. You don't even have to clear the assignment ahead of time with anyone. If you turn in an extra credit problem you write yourself, please indicate how many points you think it is worth (again up to 5 points).[3] (Of course, not all points may be awarded, regardless of how many are requested.)

In general, points are along the lines of: 1 point for working an interesting example very similar to an example in class or the notes, 2-3 points if the project amounts to locating and reporting an

[2]Examples take a while to prepare, even when you have all the data ready to go–plan on this. Strange problems crop up in analysis, particularly when using real data.

[3]If this makes you uncomfortable, don't worry about it, but it makes it easier for assigning points.

interesting result from the literature, 4-5 points for developing new ideas or projects which require a fair amount of math or programming.

Vince is the one in charge of extra credit point assignments and grading. You are welcome to talk to anyone (Hans, Derek, Vince, your mother) about it, however.

Late or missing assignments/exams

Please try to turn in your homework on time, and be present for exam days. If you are not able to get an assignment in on time, or can't turn in an exam, let Derek or Hans know (ideally ahead of time). Be aware that exams given after the scheduled day may be different than those given on the scheduled day (and may be more difficult or cover different topics).

Projects

There is no plan for a project requirement in this class. However if, as the quarter progresses, folks feel like a project would be beneficial, we can discuss such a plan.

-1.2 Rough Outline

Planning is the root of disaster. This is the best guess of when topics will be covered, however it may be tricky to keep to this schedule. It may be adjusted as the quarter progresses.

Lect.	Topic	Date
1	Introduction, matrix algebra, likelihood	April 1
2	Classical Tests	April 3
3	Ordinary linear regression	April 8
4	Linear inference	April 10
5	One way anova	April 15
6	Two-way anova, interactions	April 17
7	Ancova	April 22
8	Checking model adequacy and transforms	April 24
9	Outliers and extreme observations	April 29
10	Practice of modeling, examples	May 1
	Midterm	
11	General maximum likelihood	May 6
12	Nonlinear least squares	May 8
13	Generalized linear models 1	May 13
14	Generalized linear models 2	May 15
15	Generalized linear models Practice	May 20
16	Linear mixed effects	May 22
17	Model selection	May 27
18	Extensions, overview of other tools	May 29
19	?	June 3
20	?	June 5

-1.3 Other Resources

- Faraway, Julian. 2005. *Linear Models with R.* CRC Press.
  A very clearly written book covering material from the first half of the class–ordinary least squares, anova, ancova, and related methods.
- Faraway, Julian. 2006. *Extending the Linear Model with R.* CRC Press.
  Formerly the textbook for QERM 514, a valuable reference on generalized linear models, linear mixed effects models, and related methods.

- Conquest, Loveday. 2006. *QERM 514 Coursepacket, 2006.*
  Last year's notes from the director of QERM, Loveday Conquest. Check with Hans to get a copy.
- Pinheiro, José and Douglas Bates. 2004. *Mixed-Effects Models in S and S-PLUS.* Springer.
  A fairly extensive review of linear and nonlinear mixed effects models.
- Venables, WN and DB Ripley. 2002. *Modern Applied Statistics with S, fourth edition.* Springer.
  A wide ranging handbook of how to do applied statistics with R or S-Plus, focused on the mechanics of getting a result from R; much less theory or exposition.
- Casella, George and Roger Berger. 2001. *Statistical Inference, Second Ed.* Duxbury.
  A very rigorous introduction to statistical theory, including two chapters on regression. Limited in scope to the bare-bones of linear models (early parts of this class), but extensively detailed.
- Larsen, Richard and Morris Marx. 2001. *An Introduction to Mathematical Statistics.* Prentice Hall.
  Useful only in the early parts of the class, it covers a range of classical statistical tests including anova and linear regression. It is generally fairly rigorous but approachable.
- Kutner, Michael, Christopher Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models, fifth ed.* McGraw Hill.
  The textbook for 514 before Faraway's books came out, extensive coverage of ordinary linear models and diagnostics, some linear mixed effects; no coverage of generalized linear models, maximum likelihood, or nonlinear mixed effects.
- Dobson, Annette. 2002. *An introduction to generalized linear models.* CRC.
  This short book covers most of the topics within this class and is a superb reference (particularly, perhaps unsurprisingly, for glm).

-1.3.1 Websites

`http://students.washington.edu/nesse/qerm514/` The course website.
`http://wiki.cbr.washington.edu/qerm/index.php/QERM_514` The course Wiki.
`http://students.washington.edu/zhh/qerm514/` Last year's course website.
`http://www.r-project.org` The R-project.

Lecture 0

Matrix notation and likelihood

0.1 Main ideas

- Matrix operations: Addition, subtraction, multiplication, inverse, determinant, transpose
- Span and linear independence
- Eigenvalues, eigenvectors
- Positive (negative) definite, positive (negative) semidefinite
- Linear and quadratic equations in matrix form
- Derivatives in multiple dimensions
- Random variables
- Common distributions
- Likelihood function
- Expected value and variance
- Central limit theorem
- Cochrane's theorem

0.2 Matrix results

0.2.1 Matrix operations

For notation, when representing a matrix, a capital letter will represent a matrix, while the corresponding lower case letter with subscripts will represent elements of the matrix. Thus a marix $A = [a_{ij}]$ denotes an array of elements given by a_{ij}. The dimension of a matrix is the number of rows of the array by the number of columns, usually denoted rows $\times$ columns.

Definition 0.2.1 (Matrix addition)**.** *Let A and B be $n \times m$ matrices, and c be a constant. The addition, subtraction, and constant multiplication are defined as follows.*

- $A + B = [a_{ij} + b_{ij}]$
- $A - B = [a_{ij} - b_{ij}]$
- $cA = [ca_{ij}]$

Definition 0.2.2 (Matrix transpose)**.** *Reversing the row and column position of every element of a matrix is defined as the matrix transpose, and is denoted A' or in some texts A^T. Formally, the transpose of a matrix $A = [a_{ij}]$ is defined $A' = [a_{ji}]$. Note that if A has dimension $n \times m$ then A' will have dimension $m \times n$.*

Definition 0.2.3 (Symmetric matrix)**.** *A square $n \times n$ matrix is symmetric if $A' = A$.*

Definition 0.2.4 (Vectors)**.** *A vector is matrix which has either a row or column dimension of 1 (but not both, generally speaking), conventionally called column vectors and row vectors. For these notes, unless otherwise specified, a vector will refer to a column vector. Often vectors break from the tradition of denoting matrices with capital letters.*

Definition 0.2.5 (Vector inner and outer product)**.** *For two vectors a and b of the same length, the inner product (also called the dot product) of a and b (denoted for consistency with matrix notation either $a'b$ or $b'a$, but in some texts written $\langle a, b \rangle$) is defined here as $a'b = b'a = \sum_i a_i b_i$. The outer product is more rarely found, but is defined for two vectors of length n as the $n \times n$ matrix $ab' = [a_i b_j]$.*

Definition 0.2.6 (Matrix multiplication)**.** *Let A and B be of dimensions $a \times n$ and $n \times b$ respectively. The ordinary matrix product $AB = C$ is defined as $C = [c_{ij}] = [\sum_{k=1}^{n} a_{ik} b_{kj}]$. The dimension of the product is $a \times b$. Note that this product is only defined if A has the same column dimension as the row dimension of B, and that $AB \neq BA$ in general (even when both are defined). Matrix multiplication can be thought of as an array of inner products: the ij^{th} element of C, c_{ij} is given by the i^{th} row of A being multiplied by the j^{th} column of B.*

Definition 0.2.7 (Identity matrix)**.** *Let I_n denote an $n \times n$ matrix with a one on the main diagonal (whenever the row is the same as the column) and 0 everywhere else. I_n is frequently called the identity matrix, and when the dimension is clear from context, the n is dropped.*

Theorem 0.2.1 (Properties of the identity matrix)**.** *Let I be the identity matrix of the appropriate dimension and A be any matrix. The product of A and I always returns A, that is $AI = IA = A$. In this way, the identity matrix is akin to 1 in unidimensional numbers.*

Definition 0.2.8 (Inverses)**.** *A square $n \times n$ matrix is said to have an inverse if there is a matrix A^{-1} such that $AA^{-1} = A^{-1}A = 1$. If it exists, A^{-1} is dubbed the inverse of A. If the inverse exists, it is also unique.*

Definition 0.2.9 (Determinant)**.** *Define the minor of matrix A, denoted $A_{[-i,-j]}$ to be A removing the i^{th} row and j^{th} column. The determinant of a square $n \times n$ matrix A is denoted $|A|$ or $\det(A)$ and is defined recursively: $|A| = \sum_{j=1}^{n} (-1)^{j+1} a_{1j} |A_{[-1,-j]}|$, where the determinant of a 1×1 is the only element in it. This is not particularly efficient for computation, and in general, determinants will be computed via computer.*

Theorem 0.2.2 (Properties of the determinant)**.** *Let A and B be square $n \times n$ matrices, and c is a constant. Then*

1. $|AB| = |A||B|$
2. $|A'| = |A|$
3. $|A^{-1}| = \frac{1}{|A|}$
4. $|I| = 1$
5. $|cA| = c^n|A|$

0.2.2 Linear independence

Definition 0.2.10 (Linear combination)**.** *A vector z is a linear combination of vectors $a_1, a_2, \ldots$ is a collection of constants $c_1, c_2, \ldots$ such that $z = c_1a_1 + c_2a_2 + \cdots$.*

Definition 0.2.11 (Span)**.** *The span of a collection of vectors $a_1, a_2, \ldots$ is all the possible linear combinations of the vectors.*

Definition 0.2.12 (Linear independence)**.** *A collection of vectors $a_1, a_2, \ldots$ is said to be linearly independent if no vector a_i is in the span of the other elements. A collection which does not have this property is said to be linearly dependent.*

Definition 0.2.13 (Rank)**.** *The rank of a matrix A is the dimension of the span of the columns of A. An equivalent definition is the largest number of columns which are linearly independent.*

Theorem 0.2.3 (Row and Column rank equal)**.** *The rank of a matrix is the same as the rank of the transpose. Thus rank can be defined using the number of linearly independent rows or the dimension of the span of rows as instead of columns.*

Theorem 0.2.4 (Conditions needed for inverse)**.** *The following are equivalent for a square matrix A:*

1. *A is invertible*
2. *The rank of the square matrix A is the same as the dimension*
3. $|A| \neq 0$

0.2.3 Eigen-

Definition 0.2.14 (Eigenvalues and Eigenvectors)**.** *Any number λ and vector $v \neq 0$ are termed an eigenvalue and eigenvector of square matrix A if $Av = v\lambda$.*

Theorem 0.2.5 (Characteristic polynomial)**.** *Eigenvalues can be found by solving the polynomial equation $|A - I\lambda| = 0$, termed the characteristic equation or characteristic polynomial.*

Theorem 0.2.6 (The determinant and eigenvalues)**.** *The determinant of A is the product of the roots of the characteristic equations (the eigenvalues, including some values more than once if they are multiple roots of the characteristic equation. That is, including root λ as many times as $(x - \lambda)$ divides the characteristic polynomial). A corollary of this is that a matrix has a zero determinant if and only if it has at least one zero eigenvalue.*

Theorem 0.2.7 (Trace-determinant)**.** *For any square $n \times n$ matrix A, the trace of A is equal to the sum of the eigenvalues of A. That is*

$$tr(A) = \sum_{i=1}^{n} \lambda_i \tag{1}$$

Theorem 0.2.8 (Triangular matrices)**.** *A matrix is said to be triangular if all of the elements either $i < j$ or $i > j$ are 0. A matrix is said to be diagonal if both of these are true (equivalently, there are only elements on the diagonal, where $i = j$). The eigenvalues of a triangular matrix are given by elements of the main diagonal.*

0.2.4 Positive definite

Definition 0.2.15 (Positive definite)**.** *A square symmetric matrix A is termed positive definite if all the eigenvalues are positive. Likewise a symmetric matrix is positive semi-definite if all of the eigenvalues are non-negative. Equivalent definitions hold for negative definite and negative semi-definite.*

Theorem 0.2.9 (Determinates of PD matrices)**.** *The determinant of a positive definite matrix is positive. Likewise, the determinant of a positive semi-definite matrix is non-negative. (Note that the reverse is not true–a positive determinate is not sufficient to ensure a positive definite matrix).*

Theorem 0.2.10 (Properties of PD Matrices)**.** *A positive definite matrix A has the following properties.*

1. *$x'Ax$ is positive for all non-zero vectors x.*
2. *A can be factored into $R'R$ where R is a upper-triangular matrix with all positive eigenvalues.*

Likewise for any matrix B, $B'B$ is positive semi-definite, and positive definite if B has row rank the same as its row dimension.

0.2.5 Linear and quadratic forms

Theorem 0.2.11 (Linearity of matrix multiplication)**.** *The following properties hold true of matrix multiplication for any matricies (of appropriate dimension) A, X, Y and constant c:*

1. *$A(X + Y) = AX + AY$*
2. *$A(cX) = cAX$*

These properties are the definition of a linear function. For that reason, Ax is often said to be a linear transformation of x. Note that this is close, although not exactly the definition of linear from a high school algebra class. Multiplying out an equation like $y = Ax$ does give a collection of linear equations (for example $y_1 = a_{11}x_1 + a_{12}x_2 + \cdots$), however it does not allow for a constant unless A has a column of 1's.

Definition 0.2.16 (Quadratic form)**.** *For any square matrix A, the expression $x'Ax$ is termed a quadratic form. The rank of the quadratic form is the rank of the matrix A.*

Theorem 0.2.12 (Ellipsoids)**.** *A quadratic form equal to a constant, $y = x'Ax$ can be thought of as y as a function of x. If A is positive definite, then the x which satisfy $y = x'Ax$ for a fixed y form an ellipsoid.*

0.2.6 Derivatives

There is an amazing inconsistency in multivariable calculus notation. This probably only adds to the confusion on a topic which is generally poorly understood anyway.

Definition 0.2.17 (Differentiation with by a single variable)**.** *Suppose $Y = [y_{ij}]$ is a matrix, each element of which is potentially a function of a variable x. Then*

$$\frac{\partial Y}{\partial x} = \left[\frac{\partial y_{ij}}{\partial x}\right].$$

Definition 0.2.18 (Jacobian)**.** *Suppose $(y_1, y_2, ..., y_n)' = (f_1(x_1, x_2, ...x_n), f_2(x_1, x_2, ..., x_n), ...)'$ is a multivariate mapping from x to y. The derivative of y with respect to the vector x gives a matrix.*

$$\frac{\partial y}{\partial x} = \frac{\partial (y_1, y_2, ...)}{\partial (x_1, x_2, ...)} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

This matrix is termed the Jacobian matrix. The determinant of this matrix is termed the Jacobian determinant. Confusingly, both are sometimes referred to as "the Jacobian."[1]

Theorem 0.2.13 (Basic results for linear and quadratic forms)**.** *Let x be a vector and A be a matrix constant with respect to x. Then*

- $\frac{\partial}{\partial x} x'x = 2x'$
- $\frac{\partial}{\partial x} Ax = A$
- $\frac{\partial}{\partial x} x'Ax = x'(A' + A)$

0.3 Likelihood

0.3.1 Random variables

This development follows Casella and Berger's form, avoiding technical points of measurable sets, measurable functions, or sigma algebras. There are many texts which go into the details of these problems more thoroughly and carefully than here.

Definition 0.3.1 (Random variables)**.** *A random variable X is a map from the set of possible outcomes of an event to the real numbers. The "support" of random variable X is all the possible values X can take on (with non-zero probability). Conventionally, random variables are given a*

[1] To add to the confusion, some books use the transpose of this matrix as the Jacobian. For general interest, Wikipedia's excellent page on matrix calculus, Casella and Berger, and Mangus and Neudecker's *Matrix Differential Calculus* all seem to use definition here. Michael Perlman's notes (who taught 512 and 513 this year) uses the transpose of this matrix, as does O'Neil's *Introduction to Finite Element Methods*, which has a very clear exposition of the basic results available online. Fortunately for our purposes, the results are the same up to putting in a transpose here and there.

capital letter (which can lead to confusion, since they need not be matrices which are also often written in capitals.).[2]

Definition 0.3.2 (Cumulative probability)**.** *The cumulative density function (cdf) is the function of x which gives the probability that a random variable X is less than or equal to x. Traditionally cdfs are written with a capital letter, and when the context is not clear, the random variable under consideration is denoted with a subscript. That is*

$$F_X(x) = P(X \leq x)$$

Definition 0.3.3 (Probability density)**.** *If random variable X has a continuous support (for example, an interval of the real line), the X is called a continuous random variable. If $F_X(x)$ is also everywhere differentiable*[3]*, then the derivative of F is called the probability density function (pdf) and is usually denoted with a lower case letter corresponding to the cdfs capital. Thus the pdf of X would here be $f_X(x)$. As with the cdf, when the random variable being considered is clear from context, the subscript is generally dropped.*

Theorem 0.3.1 (pdf integration)**.** *The probability a continuous random variable takes on a value between a and b is given by the integral of the pdf from a to b.*

$$P(a \leq X \leq b) = \int_a^b f_X(t)dt$$

Definition 0.3.4 (Probability mass)**.** *If a random variable X has a discrete support (for example, the integers or the natural numbers), then it is usually called a discrete random variable. The probability of any given element in the support can be given a specific probability. The function with describes the probability of any given element is the probability mass function.*

Definition 0.3.5 (Joint density)**.** *Several random variables can have a joint density defined analogously to the univariate case. Define the joint copula*[4] *in the same manner as a cdf.*

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

The mixed partial derivative with respect to each of the x_i is the joint pdf, if it exists. Likewise, the joint pmf is the $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. Both the discrete and continuous functions are called joint densities.

Definition 0.3.6 (Conditional density)**.** *The conditional density of a random variable X conditioned on another random variable Y is their joint density over the density of Y, provided each exist.*

$$f_{X|Y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Definition 0.3.7 (Independence of random variables)**.** *Two random variables X and Y are independent if $X|Y$ has the same density as X. Independence is denoted symbolically as $X \perp\!\!\!\perp Y$.*

[2] Additionally, some random variables are traditionally written differently; ϵ_i is conventionally a normal random variable used for an error in a linear model.

[3] In practice, if X is continuous $F_X(x)$ always will be differentiable; counterexamples, such as the Cantor distribution, do exist but are rarely if ever encountered in practice.

[4] Really. That's what its called.

Definition 0.3.8 (Parameters)**.** *Often it is useful to describe families of distributions which have common properties. These distributions are indexed by parameters, generally a finite collection*[5] *of real numbers. These parameters usually describe something about the shape of the distribution of the random variable, and one of the principle tasks tasks is to find the parameters which describe the distribution from which a given data set is derived.*

0.3.2 Common distributions

Normal and related distributions

Definition 0.3.9 (Normal distribution)**.** *A normal or Gaussian density is a continuous family of pdfs with two parameters, μ and σ^2, which has the form*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Conventionally the density where, $\mu = 0$ and $\sigma^2 = 1$ is called the standard normal density. Since the normal density comes up fairly frequently, it is often written $\phi_{\mu,\sigma^2}(x)$, or for the standard normal, $\phi(x)$. The standard normal cdf function is likewise often written $\Phi(x)$. When describing the random variable, the normal is often written $N(\mu, \sigma^2)$. (Note that some sources, including `R`*, prefer to describe the normal using σ, rather than σ^2, but this is purely a matter of convention.)*

Theorem 0.3.2 (Linear transformations of the normal)**.** *Let $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$, and let $X \perp\!\!\!\perp Y$. Let a and b be univariate constants. Then the following are ture.*

1. $aX + b \sim N(a\mu_x + b, a^2\sigma_x^2)$
2. $X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$
3. $X - Y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$

Definition 0.3.10 (Multivariate normal)**.** *The multivariate normal is, perhaps unsurprisingly, the multivariate extension of the normal. A collection of n random variables have a multivarite normal distribution, with mean vector μ and covariance matrix Σ if the pdf is as shown below.*

$$f_X(x) = f_{(X_1,X_2,\ldots,X_n)'}(x_1, x_2, \ldots, x_n) = \frac{1}{|2\pi|^{n/2}\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

Theorem 0.3.3 (Linear transformations of the multivariate normal)**.** *Let A and b be a matrix and vector respectively of dimension $k \times n$ and k. If $X \sim N_n(\mu, \Sigma)$, then $AX + b \sim N_k(A\mu + b, A\Sigma A')$.*

Definition 0.3.11 (Chi-squared distribution)**.** *A random variable S is said to have a χ_n^2 distribution if it can be written as the sum of n independent $N(0,1)$ random variables squared. That is if $X_i \overset{iid}{\sim} N(0,1)$ then $\sum_{i=1}^n X_i^2 \sim \chi_n^2$. Often the chi-squared parameter (here n) is called "degrees of freedom."*

Definition 0.3.12 (t distribution)**.** *A standard normal distribution over the square root of an independent χ_n^2 is said to have a Student's t distribution. Note that the n is inherited from the χ_n^2,*

[5] A family which requires an infinite number of parameters to describe is called non-parametric.

so the t distribution has a single parameter, n. Like for the χ^2_n, this parameter is often called the degrees of freedom. That is, if $X \sim N(0,1)$ and $S^2 \sim \chi^2_n$ such that $X \perp\!\!\!\perp S^2$ then

$$\frac{X}{\sqrt{\frac{1}{n}S^2}} \sim t_n \tag{2}$$

Definition 0.3.13 (F-distribution)**.** *The ratio of two independent χ^2 random variables, on n and m degrees of freedom, each divided by their degrees of freedom, is distributed as $F_{n,m}$.*

$$\frac{\chi^2_n/n}{\chi^2_m/m} \sim F_{n,m}$$

Common discrete distributions

Definition 0.3.14 (Bernoulli)**.** *A random variable X is said to be Bernoulli(p) if it takes on the value 1 with probability p and the value 0 with probability $1-p$.*

Examples of Bernoulli random variables include anything with only two possible outcomes: (potentially biased) coin tosses, the Mariners winning ($X=1$) or losing ($X=0$) a particular game, etc. The equation which describes the probability mass function for a Bernoulli trial is not very illuminating:

$$f_{Bernoulli}(x) = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases} = p^x(1-p)^{1-x} \tag{3}$$

Note that the output of the pmf is the *probability* and input is the value of the random variable.

Definition 0.3.15 (Binomial distribution)**.** *The binomial distribution with parameters n and p is the sum of n independent Bernoulli random variables, each with parameter p. The pmf of the binomial is*

$$f_{binom}(x) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{4}$$

Binomial events come up frequently as the sum of independent Bernoulli random variables. For instance, if a random variable X represents a coin-toss with heads being $X=1$, then the number of heads H in ten coin tosses X_i is just $H = \sum_{i=1}^{10} X_i$. Here H will be distributed binomial(10,p), where p is the probability of getting a heads on one toss (0.5 for a fair coin).

A multinomial is a multivariate generalization of the binomial. Whereas the binomial was the sum of random variables, each of which had two possible outcomes, the multinomial is the sum of random vectors each of which has k possible outcomes.

Definition 0.3.16 (Multinomial)**.** *Represent an event X_i with k possible outcomes as a vector of length k. If the outcome of the event is i, the random vector is all zeros except for a 1 in the ith position (i can be 1 to k). Further let the probability of the ith outcome occurring be p_i. The multinomial$(N; p_1, p_2, \ldots, p_k)$ is the sum of N such independent events.*

Multinomial distributions are heavily used in ecological experiments, such as mark-recapture. Often the description of the multinomial is of a collection of k boxes numbered 1 though k and N balls. If each ball is tossed and has probability p_i of landing in box i, a multinomial random vector is the counts in each box after all the balls have been tossed. Note that $\sum p_i = 1$, and that the maximum count in any box is N.

Name	Parameters	Support	pmf	Description
Bernoulli	p	$\{0,1\}$	$f_X(x) = p^x(1-p)^{1-x}$	The Bernoulli distribution describes the probability of a random variable X being either a 0 or a 1, where p describes the probability of getting a 1.
Binomial	n,p	$\{0,1,\dots,n\}$	$f_X(x) = \binom{n}{x}p^x(1-p)^{n-x}$	The binomial distribution describes the number of successes in a series of n independent trials, where p is the probability of success of one trial.
Poisson	λ	$\{0,1,\dots\}$	$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$	The Poisson model is often described as the count of things arriving at a constant rate.
Negative Binomial	r,p	$\{0,1,\dots\}$	$f_X(x) = \binom{r+x-1}{x}p^r(1-p)^x$	The description often given is the number of failures to the rth success. Some sources describe this distribution as the Pascal distribution, and describe a generalization of this as the negative binomial.
Discrete Uniform	n	$\{1,2,\dots,n\}$	$f_X(x) = \frac{1}{n}$	The discrete uniform describes the probability of n possible outcomes, each of which has equal probability.
Geometric	p	$\{1,2,\dots\}$	$f_X(x) = p(1-p)^{x-1}$	The geometric distribution describes the probability of one success (which occurs with probability p) at the end of a string of independent trials. It is a special case of the negative binomial.

Table 1: Brief summaries of commonly used univariate discrete distributions. Not all of these are used in the course, and of course, this list is not complete (in fact, no list possibly could be–there are infinitely many distributions).

The multinomial has a pmf which assigns a probability to each possible vector of outcomes $x = (x_1, x_2, \dots, x_k)'$.

$$f_{multinom.}(x) = \frac{N!}{x_1!x_2!\cdots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k} \tag{5}$$

Another discrete random variable, this one without an upper bound on the support, is the Poisson distribution.[6]

Definition 0.3.17 (Poisson)**.** *A random variable X is Poisson distributed with parameter λ if X is the count of the number of events occurring in a specific interval of time, when events occur as a "Poisson process." The Poisson process has five requirements:*

- *X starts with 0*
- *The count over disjoint time intervals are independent*
- *Counts over intervals of equal length are identically distributed*
- $\lim_{t\to 0} P(X_t = 1)/t = \lambda$
- $\lim_{t\to 0} P(X_t > 1)/t = 0$

[6]Named after the French mathematician Siméon Denis Poisson. Note that *poisson* also means fish in French leading to no end of statistical puns, few if any of which are funny.

Many things are modeled with a Poisson distribution. For instance, the number of people arriving at a theater might be a Poisson process, in which case the number of people who arrive at the theater in a 10 minute period is a Poisson random variable.[7] Another example is the counts of a radioactive element decaying, at least over a short period of time relative to the element's half life.

It is a (non-trivial) theorem that the Poisson postulates give rise to the pmf for the Poisson distribution shown below.

$$f_{Pois.}(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{6}$$

Common continuous distributions

There are no continuous distributions which are used in the course which are not based on the normal distribution. Like the discrete case, however, there are infinitely many possible distributions out there which might be encountered outside of the course. The uniform(a, b) distribution gives equal probability to any value between a and b (it is commonly what people think of as a "random" number). A generalization of the uniform, the beta distribution, ranges between zero and one and has a wide range of possible shapes. A gamma distribution, which has support from 0 to infinity, takes on a wide range of shapes as well. A special case of the gamma is the exponential distribution, which is also the distribution of the time between Poisson process events.

0.3.3 Likelihood functions

Definition 0.3.18 (Likelihood function)**.** *Data are generally modeled as random variables X with distributions $f_X(x|\theta)$ with one or more unknown parameters θ. Once the data are observed, the distribution can be thought of as a function of θ. This is known as the likelihood function. Note that although it shares the form of a probability distribution, as a function of θ most often f as a function of the parameters is not interpretable as probability distributions.*

0.3.4 Expected value and variance

Definition 0.3.19 (Expected value)**.** *The expected value EX of a random variable X is defined as*

$$EX = \int_{-\infty}^{\infty} t f_X(t) dt$$

for continuous distributions, and

$$EX = \sum_{x \in Support} x f(x)$$

for discrete random variables. Note that a expected value need not exist for any specific distribution. More generally, the expected value of a function $g(X)$ is

$$Eg(X) = \int_{-\infty}^{\infty} g(t) f_X(t) dt$$

[7] This is a classic example, but would be incorrect to apply to reality here. Theater-goers are likely to arrive in groups, which violates the last of the Poisson postulates. Furthermore, the count in the last 10 minutes before the show starts is likely to be differently distributed from the 10 minute interval an hour before the show starts, violating the third assumption.

Definition 0.3.20 (Variance and standard deviation)**.** *The variance of a random variable X is defined as $VarX = EX^2 - (EX)^2$, provided the expectations exist. The standard deviation is the $\sqrt{VarX}$.*

0.3.5 Key results

Theorem 0.3.4 (Law of large numbers)**.** *For a collection of random variables X_i which are independent and identically distributed with a probability density which has expected value μ and bounded variance (that is, $VarX < \infty$), then the mean of the X_i, $\bar{X} = \frac{1}{n}\sum X_i$, converges to the mean of the distribution μ.*

0.3.6 Central limit theorem

The central limit theorem is a fairly technical result, for which there is a somewhat inaccurate approximation which is useful.

Theorem 0.3.5 (Central limit theorem)**.** *Suppose X_i is a collection of independent identically distributed random variables*[8] *with mean μ and variance $\sigma^2 < \infty$. Then*

$$\sqrt{n}(\bar{X} - \mu) \to N(0, \sigma^2)$$

A useful way to think about this, however, $\bar{X} \dot{\sim} N(\mu, \sigma^2/n)$. This can be misleading, however, since it is only true as $n \to \infty$, in which case this becomes a degenerate distribution.

Theorem 0.3.6 (Cochrane)**.** *Suppose $X'X$ is distributed χ^2_n and further $X'X = X'Q_1X + X'Q_2X + \cdots + X'Q_kX$, where each Q_i is a positive semidefinite matrix with rank r_i. Any one of the following statements implies the other two:*

1. $\sum r_i = n$
2. $X'Q_iX \sim \chi^2_{r_i}$ *for all i*
3. $X'Q_iX$ *are all mutually independent*

0.4 Examples

0.4.1 Handy R commands

The most useful R commands are for finding functions or finding out more about a function. To get help on a function (introduced in the text or found elsewhere) use either a questionmark before the command, or `help()`. For instance, to get help on the `lm()` command (which will be used extensively), try either of the commands below.

```
> ?lm
> help(lm)
```

[8] This can actually be generalized as well.

In these notes, functions will always be presented with parentheses following (such as `lm()`).

In many cases, the function name for a particular function will be unknown. To find a function, use the `help.search()` command. For instance, to find functions which have to do with box plots, try

```
> help.search("box plot"')
```

The results of the `help.search()` command will be a list of function names, followed by a library name in parentheses. To load one of these libraries, either use the drop-down menu "packages" or the `library()` command. For instance, to load the MASS library (which has many useful functions), use

```
> library(MASS}
```

The default library (which does not need to be loaded) is `stats`. In these notes, the need to load additional libraries will be indicated.

Moving on to actual functions, the prompt of `R` can be used as a calculator. For instance

```
> 5+2^3-3
 [1] 10
```

However `R` is much more powerful. Variables can be created using the `<-` command[9].

```
> x<-3
> x+1
 [1] 4
```

Likewise vectors can be entered in `R` using the concatenate command `c()`. For example,

```
> x<-c(1,2,3)
> x
 [1] 1 2 3
```

Specific elements of a list can be called using square brackets. Likewise, an element can be replaced by assignment.

```
> x[2]
 [1] 2
> x[2] <- 5
> x
 [1] 1 5 3
```

Likewise, multiple elements of the vector can be called by putting another vector in the square brackets. Thus calling the first and third elements of `x`

```
> x[c(1,3)]
 [1] 1 3
```

[9]It is also possible to use the `=` assignment, however the arrow is a bit clearer and is recommended by the `R` development team.

It is common to need to create vectors which are sequences of numbers or repeated collections of numbers. A sequence can either be created using the `seq()` command or a colon, `:`. Vectors with repeated entries are created using the `rep()` command.

```
> x <- seq(1,5,0.5)    # Creates a sequence from 1 to 5 (inclusive) incrementing by 0.5
> x
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> x <- 1:10            #Makes a sequence from 1 to 10 by 1, eqivelent to seq(1,10,1)
> x
 [1]  1  2  3  4  5  6  7  8  9 10
> x <- rep(1,10)       # makes a vector of length 10 of all 1s
> x
 [1] 1 1 1 1 1 1 1 1 1 1
```

Often it is useful to use datasets created elsewhere. For this purpose, there are three commands which are very useful: `read.table()`, which reads tab-delimited tables, `read.csv()` which reads comma seperated variables, and `scan()` which is an all-purpose text reading command. The file can be stored on your hard drive, or `R` can access the internet to get it. The `read.table` and `read.csv` commands both form data frames in `R` (see below).

```
> x <- read.table(file='C:\\Documents and Settings\\Hans\\My Documents\\514\\filename.txt',header=T)
> y <- read.table(file='http://students.washington.edu/nesse/qerm514/data/kurilTerMol.txt',header=T)
```

Note that if the file path for a local file is copied and pasted into `R`, all of the backslashes must be doubled.

0.4.2 Data frames

One of the useful data structures in `R` is the data frame. A data frame is a collection of vectors of equal length, with a name for each. To create a data frame, use the `data.frame` command.

```
> a1<-c(1,4,6,4,1)
> a2<-c('z','z','w','w','w')
> a3<-c(1,0,0,1,-1)
> data.frame(a1,a2,a3)
  a1 a2 a3
1  1  z  1
2  4  z  0
3  6  w  0
4  4  w  1
5  1  w -1
```

Like any other object in `R`, a data frame can be stored as a variable. The dollar sign, $, is used to access vectors within the data frame.

```
> x <- data.frame(a1,a2,a3)
> x$a1
 [1] 1 4 6 4 1
```

The data frame names can also be displayed (or modified) using the `names()` command.

```
> names(x)
 [1] "a1" "a2" "a3"
> names(x) <- c('a','b','c')
```

0.4.3 Entering a matrix in R

Suppose we want to enter the following matrix in R:

$$\begin{pmatrix} 1 & 2 & 6 \\ 2 & 0 & 5 \\ 0 & -1 & 1.3 \end{pmatrix}$$

```
> x<-matrix(c(1,2,0,2,0,-1,6,5,1.3),3,3)
```

Note that R fills up the matrix by column, unless otherwise specified.

It is also possible to coerce a rectangular object, such as a dataframe, to be a matrix. Suppose y is already a data frame here. The `as.matrix` command can respecify the type to be a matrix.

```
> y<-as.matrix(y)
```

To recall an item from a matrix, use square brackets. For instance, to call the 5 from the matrix x created above, note that it is in the 2nd row and 3rd column.

```
> x[2,3]
 [1] 5
```

0.4.4 Matrix operations

The matrices entered in R can be manipulated using the `+`, `-`, `%*%` commands for addition, subtraction, and matrix multiplication. Transpose is `t()`, determininate `det()`, and inverse is `solve()`. **Note:** Often the"intuitive" command (such as `x^(-1)` for inverse or `*` for multiplication) will not give any errors; they are also commands. They are not, however, the commands you'd expect (`x^(-1)` for instance, inverts every element of x, not the matrix.). Most commands enter the matrix element-wise.

0.4.5 Distributions in R

There are many probability density functions, cumulative density functions, and random number generators available in R. Some of the common ones are shown in table 0.4.5.

For example, 10 independent draws from a uniform(0,1) random variable can use the `runif` command.

```
> runif(10,0,1)
 [1] 0.08789804 0.17243325 0.05399578 0.79523893 0.37975959 0.60435588
 [7] 0.79250848 0.21006469 0.47955394 0.80843348
```

	Random number	Density	cdf	Quantile
Normal	`rnorm`	`dnorm`	`pnorm`	`qnorm`
Uniform	`runif`	`dunif`	`punif`	`qunif`
χ^2	`rchisq`	`dchisq`	`pchisq`	`qchisq`
t	`rt`	`dt`	`pt`	`qt`
F	`rf`	`df`	`pf`	`qf`
Gamma	`rgamma`	`dgamma`	`pgamma`	`qgamma`
Exponential	`rexp`	`dexp`	`pexp`	`qexp`
Beta	`rbeta`	`dbeta`	`pbeta`	`qbeta`
Binomial	`rbinom`	`dbinom`	`pbinom`	`qbinom`
Poisson	`rpois`	`dpois`	`ppois`	`qpois`

Table 2: Functions for random variables built into `R`. Random gives random numbers from the distribution, Density returns the pdf or pmf value of the argument, cdf gives the probability of a value $\leq$ than the argument, while quantile returns the value which is at the quantile given in the argument.

Lecture 1

Introduction

The field of statistics is very large. This is a course in a very limited subset of those statistics, dealing with classical linear models and a few generalizations. It is ecologically focused in the sense that the topics covered are some of methods an ecologist is likely to either need or come across in the literature. The examples are not always ecological.

1.1 Questions

Roughly speaking, there are four types of statistical questions this course is aimed at answering, at least to some degree.

Estimate	What parameters, in a particular model, best fit the data?
Inference	How certain are those estimates and what can be said about them?
Adequacy	Is the model probably the right choice?
Prediction	When can predictions be made for new observations?

The amount of detail on each of these topics varies. And it is worth noting that many (useful) questions are not included. Questions, for example, related to the design of an experiment (how to collect data, and how much) are not included in the course. That is not to say that other questions are not common in ecology; only that time is limited.

1.2 Data

The sort of data which this course examines is a collection of observations. Each observation is made up of one or more predictor variables, and one response variable.[1] For this course, the response

[1]The terms "predictor" and "response" go by many other names in other sources. A predictor variable is also sometimes called a covariate, regressors, input, explanatory variable, or (confusingly) either a dependent or independent variable. The response variable is sometimes also called the output, endpoint, or the dependent/independent variable. The last of these, dependent and independent variables is perhaps the worse. Which variable is called independent and which dependent changes depending on the source.

of a particular observation i will always be a single value denoted with a y_i.[2] The predictor or predictors will usually be combined into a matrix X, however sometimes particular predictors j within an observation i will be written separately, as x_{ij}.

The data analyzed in this course is specific in several ways.

- It is composed of observations, which are one or more predictors and a single response.
- The responses, given the predictors, are independent.
- The differences between the predictions and the response are distributed according to one of a few distributions, described later.
- The response for each observation is composed of one value.

The types of variables which can compose the observations are also somewhat limited. Variables can come in many forms. The three types which will be used in this course are below.

Continuous	Variables which can take on any value on a interval (potentially the whole real line).
Categorical	Variables which can take on a finite collection of values, termed "levels." Special types of categorical variables include
Binary	Categorical variables with only two levels.
Ordinal	Categorical data which also has a natural ordering to the levels (such as "low, medium, and high" for instance).
Count	Variables which can take on only integer values.

Of course, there are many other variable types which are not being covered. Circular variables (which have some kind of wrapping property) for instance, such as month or cardinal direction, do crop up now and again in ecological research but are ignored here.

1.3 Models

In statistics, a model is a function (usually with one or more free parameters), which describe the observations. For the data that this course covers, the form of the equation is $y_i = g(X_i, \epsilon_i)$, where y_i is the response for the observation i, X_i is the predictors for the observation i, and ϵ_i is one or more random variables.

1.3.1 Linear models

Classical linear models

The first half of the class is built around a single model.

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)}}_{X_i\beta} + \epsilon_i \tag{1.1}$$

[2] There is an extensive literature on generalizing this, so that there can be multiple responses within an observation; this generalization falls under the rubric of "general linear models," not to be confused with "generalized linear models" which *are* a topic in this course.

Here the β_j are free parameters, and the ϵ_i are each independent and normally distributed. This is the model for ordinary linear regression, analysis of variance (anova), and analysis of covariance (ancova). These are all combined under the rubric of multiple regressions.[3]

The course will often take advantage of matrix notation and write lengthy expressions concisely. Here $\beta_0+\beta_1 x_{i1}+\cdots+\beta_{p-1}x_{i(p-1)}$ can be written as the matrix product of a row-vector of predictors X_i times a column vector of coefficients β.

Generalized linear models

One generalization of the linear model is the (cleverly named) generalized linear model. The generalized linear model has two additional components over and above the classical model. It has an link function g, and the error can be one of a collection of distributions, rather than normally distributed as in the classical case. These are subject to the following limitations.

- The link function g must be invertible (and in practice will be one of a small collection of specified functions)
- The link and error must be chosen such that the expected value of the response must be $\mathrm{E}Y_i = g^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i(p-1)}$.

Linear mixed effects models

Some data have a hierarchical structure, and the questions of interest are those at the higher level. For instance, in a forestry experiment which involves many experimental plots, the experimenter is probably not interested in the effect of any one plot, but rather the mean and variation of an effect averaged over all the plots. Linear mixed effects are a class of models which can capture this. The structure of the model is that some (or all) of the parameters used to make the model predictions are random variables. Formally this is written (using matrix notation)

$$y = X\beta + Zb + \epsilon \tag{1.2}$$

Here β is a vector of coefficients β_0 through β_{p-1} and X is a matrix of predictors. The other matrix Z is also a matrix of predictors, but the coefficients b are normally distributed random variables rather than being free parameters. The parameters of interest to the researcher, however, are not the specific values of b which fit the data, but the parameters of the distribution which gave rise to b (along with an interest in the values of β). The ϵ is, like in the classical linear model, a normally distributed random variable.

1.3.2 Non-linear models

Maximum likelihood estimates

The title here is misleading, since almost all of the parameter fitting, in the linear and nonlinear sections alike, using maximum likelihood methods. However most of the time, the process is hidden so as to work efficiently. In some cases, this can not be done reasonably well and the likelihood must be explicitly maximized.

[3] Not to be confused with multiple anova, or manova, which has a multivariate response y (rather than univariate as here) for each observation.

Almost any model $y_i = g(X_i, \epsilon_i)$ (including non-normal errors) can be fit using maximum likelihood methods.[4]

Nonlinear least squares

One commonly-used method for fitting parameters θ to a user specified function $g(X, \theta)$ is nonlinear least squares (nls). Strictly speaking, the nls procedure does not assume a specific model. Rather it numerically finds the values for θ such that $\sum_i (y_i - g(X_i, \theta))^2$ is minimized. This method does have some nice properties for the model $y_i = f(X_i) + \epsilon_i$ where ϵ_i is a normal random variable with mean 0.

Nonlinear mixed effects models

As the name suggests, it is a nonlinear extension of linear mixed effects. Like the linear mixed effects, observations are often from a heirarchical arrangement (that is, data fall into natural groups, but the parameters of particular groups are not of interest). Parameters for a nonlinear function are assumed to come from another distribution, the parameters of which are of interest. The model, for a nonlinear (but differentiable) function g, observed responses y, random parameters θ, and normal error ϵ. The response Y_{ij} (observation i in group j is given by

$$Y_{ij} = g(X_{ij}, \theta_{ij}) + \epsilon_{ij} \tag{1.3}$$

The θ_{ij} however, come are equal to

$$\theta_{ij} = X_{ij}\beta + Z_{ij}b_j \tag{1.4}$$

and as with the linear models, $b_j \sim N(0, \Psi)$.

[4] There are some limitations of course, the MLE may not exist in some cases, or may be ill defined.

Lecture 2

Classical tests

2.1 Main ideas

- Exploratory data analysis plotting
- Null and alternative hypotheses
- p-value
- t-test
- χ^2 goodness of fit
- Randomization test

2.2 Exploratory data analysis

Describing data can be a difficult topic. What form of exploratory analysis is performed is dependent on the form of the data, and what the researcher suspects is interesting about it.

2.2.1 Summary statistics

Definition 2.2.1 (Mean, median)**.** *The mean of a collection of points $\{x_1, x_2, \ldots, x_n\}$ is found by the aritmatic average, and usually denoted $\bar{X}$.*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The median of a collection of data is found by first ordering the data. Let $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ denote the points put in numerical order. The median, usually denoted by m is defined slightly differently depending on whether n is even or odd.

$$m = \begin{cases} \frac{x_{(n/2)} + x_{(n/2)+1}}{2} & n \text{ even} \\ n_{(\frac{n+1}{2})} & n \text{ odd} \end{cases}$$

Notably, these do not estimate the same thing in general. The mean $\bar{X}$ estimates the mean of the distribution, when it exists, while the median of the data estimates the median of the distribution (which always exists).[1]

The median, already described, could be called the middle of the data. It is constructed such that half of the data is smaller than the median, and half of the data is larger. A reasonable extension of this is to define quantiles–numbers such that q percent of the data are below the number.

Definition 2.2.2 (Quantiles and IQR)**.** *The pth quantile of a dataset $x_1, \ldots, x_n$ is a number q_p such that p percent of the data are less than the value. There are various ways of specifying a number when q_n is not exactly on a datum; (*`R` *has nine options in the* `quantile` *function), but the principle is the same. A specific case of quantiles is quartiles–the first, second, and third quartiles are defined as the 0.25, 0.5, and 0.75 quantiles. The inner-quartile range (IQR) is the difference of the .75th quantile and the .25th, $IQR = q_{0.75} - q_{0.25}$.*

A good place to start with a collection of univariate data, that is just a list of numbers, is to calculate some measure of the center of the data, and some measure of how varied the data are. Commonly, the mean of the data is used to estimate the center, while the median could also be used. The variance or standard deviation can estimate the variability of the data, alternatively the interquartile range (IQR) could be used.

The motivation for using the median and IQR instead of the of the mean and standard deviation is often the expectation of outliers. An outliers is technically defined as a number which is greater than 1.5 times the interquartile range away from the nearest quantile. However in a more colloquial sense, an outlier is usually thought of as a result which came about by a different process (such as an observation error). While one approach is to throw out outliers, use of the so-called "robust" statistics (the median and IQR) allows the inclusion of all the data.

Additionally the difference between the median and mean can give an indication of the skew of the data.

2.2.2 Box plot

The boxplot is a recent invention–1977 by John Tukey, according to Wikipedia. It is most often used when data come from several known groups. That is, each data point has a group associated with it. Data points to be plotted would look like (a, x_1), (a, x_2), (b, x_3),..., where a and b represent groups (for instance, treatment and control).

A boxplot, also called a box and whisker plot, displays multiple data for each group, see figure 2.1. The median and first and third quantiles are simply summary statistics described above. The distinction between outliers and non-outliers is fairly arbitrary, but in `R` (and elsewhere), an outlier is defined as an observation greater than 1.5 times the IQR away from the nearest quantile.

2.2.3 Mosaic plots

Introduced in the early 1980s, the mosaic plot is a slick way of displaying count responses which have multiple categorical predictors (said another way, a contingency table). For the sake of brevity,

[1] This might seem odd the first time reading it; the *data* always have a mean. It is always possible to take the average of a collection of data. Many distributions also have means, formally defined for continuous distributions as $\int xf(x)dx$ where $f(x)$ is the pdf of the distribution.

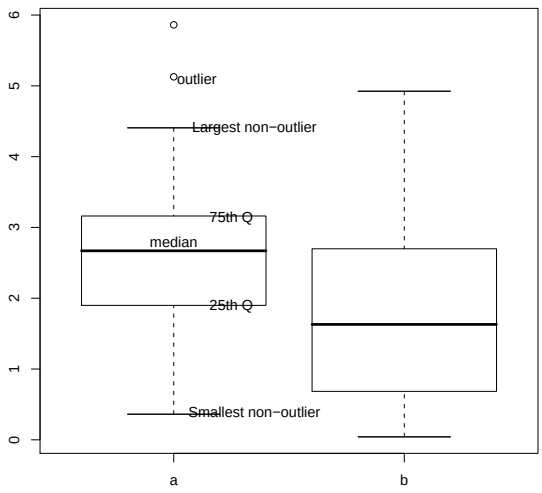


Figure 2.1: A boxplot of two data sets,"a" and "b", with the parts of the boxplot labeled on the first group.

only a two by two table is shown below. Using the NISP data from the archaeology example in section 2.8.2, suppose two sites have bone counts (NISP) from two different species (seals and sea lions). For instance

	Site 1	Site 2
sea lion	266	419
seal	516	558

We might be concerned, for instance, that the species makeup at the two sites is different. To assess this visually a mosaic plot can be used. See the results displayed in figure 2.2. The width of the two columns indicates the number of observations in each site (thus site 2 has more bones total than site 1), while the height of the two boxes indicates the proportion of bones from each species at each site.

Mosaic plots can be extended to larger data sets, comparing many levels within a categorical predictor, or even many categorical predictors.

2.3 Hypothesis testing

Data are rarely interesting for their own sake. The results which make research interesting is the interpretation. Most of statistical hypothesis testing has been built around trying to rule out "randomness" giving rise to a collection of data. The null model is an attempt to capture what randomness might be at work; sometimes this is meaningful (and sometimes it is not).

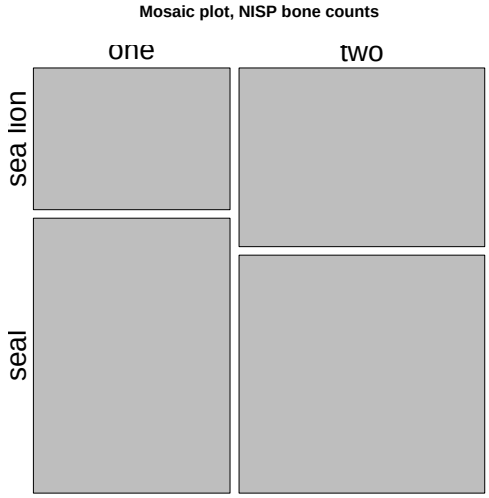


Figure 2.2: A mosaic plot of the NISP bone counts from the archaeology example in section 2.8.2.

The form that a null model takes is generally left to an experimenter. For example, take the data shown in figure 2.1: The data are responses from two different groups. The null hypothesis could be very general, such as the hypothesis came from the same distribution. Alternatively, if the origins of the data suggest that they necessarily are normally distributed, it might be reasonable to test whether they came from the same normal distribution. Further, perhaps if the process indicates that the two groups not only both come from a normal distribution, but should have the same variance, the null hypothesis might be equality of the means of distributions underlying the data.

Since multiple null hypotheses are possible for any given dataset, it is good practice to explicitly state the null hypothesis under consideration.

Alternative hypotheses

If there is a null hypothesis, and it may be rejected, there must be an alternative. The exact specification of the alternative hypothesis is left to the experimenter. In general, the alternative is fairly intuitive from the problem; if the data are modeled as normal with mean $\mu = \mu_0$ under the null, for instance, the alternative might be normal with mean $\mu \neq \mu_0$ (called a two sided alternative) or $\mu > \mu_0$ (called a one-sided alternative). In some cases (particularly in non-parametric tests), the null is taken to be the negation of the null hypothesis.

Many of the inference techniques can be thought of as comparing two models–usually described as the null and alternative. In most cases in these notes, the deeper philosophy behind a development of hypothesis testing is ignored.

Equal Variance	The classical t-test. This test tests the null hypothesis that two groups of observations have the same mean and the same variance, against the alternative that they have different means and/or variance. The formal statement of the null hypothesis is that both groups of observations come from the same $N(\mu, \sigma^2)$ distribution.
Welches t-test	A variant of the equal variance t-test, tests the hypothesis of equality of means between two groups, without assuming the groups have equal variance. Formally, if the first group is iid $N(\mu_1, \sigma_1^2)$ and the second group of observations is iid $N(\mu_2, \sigma_2^2)$, the null hypothesis is that $\mu_1 = \mu_2$.
One sample	A test to determine if a single data set has a specified mean. More formally, the null hypothesis is that the data are distributed $N(\mu, \sigma^2)$, with μ specified and σ^2 unknown.
Paired t-test	A test on data which naturally come in pairs of observations, testing whether the difference of each pair is close to zero. Formally, the null hypothesis specifies that each pair i of points comes from the same $N(\mu_i, \sigma^2)$ distribution. While each observation (in each pair) have a common variance σ^2 shared by all pairs, each pair of observations can have a different mean μ_i.

Table 2.1: Common types of t-tests.

2.3.1 p-value

In a hypothesis test, the p-value is the probability that the data would have come up in the way they did, or "worse" (what it means to be worse has to be specified by the alternative hypothesis), if the null hypothesis were true. Thus a very low p-value is an indication of something other than the null model at work.

Note that the p-value should not be interpreted as the probability that the alternative hypothesis is true; sample size plays a large role in determining the p-value as well as any actual difference.

2.4 t-test

There is no single "t-test", rather there are several tests based on the t distribution (see definition 0.3.12 for more information on the t-distribution) which get called t-tests. Differences arise due to different specification of the null hypothesis, and several types are summarized in table 2.4. No matter the type of t-test, each can be performed in as one-sided or two-sided alternative. In a two-sided test, the equality under consideration is tested against inequality. A one-sided test examines less-than or equal to against greater than. See examples for illustration of each of these.

2.5 χ^2 test

Just as the t-test refers to any test using a t distribution as the null hypothesis, the chi-squared test is any test which uses the chi-squared[2] distribution (also written χ^2 distribution) as the null hypothesis. Here we'll examine only two subtly different tests based on the χ^2 distribution.

The first such test examines whether data which come from various groups fall into those groups with prescribed probabilities. Picture the setup for this test as a collection of boxes, each with a prescribed probability. In an experiment, n balls are thrown into the boxes randomly. The chi-squared test of this form tests the hypothesis that the balls fell into the boxes with the prescribed probabilities, against the alternative that they fell into the boxes according to some other box probabilities. The null hypothesis for the chi-squared test here is that the observations are distributed multinomial with the specified vector of probabilities.

The second form of the chi squared test is similar to the first. In this test, two groups of balls (not necessarily equal sized) are compared, to determine if they fell into boxes with the same box box probabilities. The null hypothesis is that each set of observations is distributed multinomial with the same vector of probabilities; the alternative hypothesis is that the probabilities are different for each group.

In either case, the general form of the statistic being calculated is

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2.1}$$

Here O_i is observed counts, E_i is expected counts. For a test against a known set of probabilities (such as the Benford's law example in section 2.8.1 below), the expected count $E_i = np_i$ (n is the total number of observations and p_i is the probability for that cell. To test whether two datasets come from the same distribution, define R_i and C_j to be the row and column totals from row i and column j respectively. The expected counts in cell ij is $E_{ij} = R_i C_j / n$.

Yates' Correction

Since the χ^2 test relies on a continuous approximation of discrete data, the approximation for small data sets can be poor. To limit this effect, at least in part, Yates' correction is used by default for certain tests in `R`. Yates' correction amounts to an adjustment of the χ^2 formula, equation 2.1. The Yates corrected form is

$$\sum_{i=1}^{n} \frac{(O_i - E_i - 0.5)^2}{E_i} \tag{2.2}$$

2.6 Randomization/permutation tests

A randomization or permutation test is a method of determining if two or more sets of data come the same distribution, without necessarily specifying what that distribution is.[3] More specifically, a

[2] Although the temptation is to pronounce the test a "chee" squared test, the actual pronunciation is more like "kai" squared test. The Greek letter χ is pronounced like the first syllable of the word kayak.

[3] Some sources make a big distinction between these–and they are technically different: A randomization test generates a null distribution by generating the groups randomly, whereas a permutation test uses all possible groupings. However the distinction is not immediately relevant for any analytic purpose in these notes, so it is glossed over.

permutation test or randomization test examines whether a specific statistic (the mean, for example) is different significantly different.

The formulation of the test is thus: suppose the data x exist in two groups, x_{ai} and x_{bi}. To test if the data in group a came from the same distribution as group b (without specifying a distribution), we need a statistic $m(x)$ (most often $m(x)$ is just the mean, but just about any other statistic can be used). The observed difference $m(x_a) - m(x_b)$ is compared against differences $m(x_q) - m(x_r)$, where q and r are randomly assigned groups (of equal size). The steps of a randomization test are

1. For some estimator[4] m, calculate $m(x_a) - m(x_b)$.
2. Randomly reassign observations into groups q and r (where q and r have the same size as a and b).
3. Recalculate the difference $m(x_q) - m(x_r)$
4. Repeat steps 2 and 3 many thousands of times
5. Determine how frequently the actual difference, $m(x_a) - m(x_b)$ was more extreme (either larger or smaller) than the randomly generated differences. That percentage becomes the p value for the statistic.

Since the randomization test can be done for a wide range of statistics, it is a versatile means of comparison. Most often, the question of interest is a difference of means, however any statistic could be used.[5]

2.7 R functions

	Description
`mean()`	Finds the mean of the argument
`median()`	Median of the argument
`var()`	Finds the unbiased variance of the argument
`quantile()`	Finds the specified quantile
`IQR()`	Inner quartile range
`boxplot()`	Plots a boxplot of the data
`plot()`	Can be used for all sorts of plotting, here used for Cartesian plots
`t.test()`	t-test
`chisq.test()`	χ^2 test for count data
`for()`	Starts a for loop

[4] An estimator is a function of the data.

[5] And in fact, I know of no reason why the difference must be used, and not say, the ratio.

d	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Table 2.2: The probabilities of getting digit d as the first digit of an observation, according to Benford's law.

First digit	1	2	3	4	5	6	7	8	9
Count of nations	57	31	28	14	11	11	9	10	12

Table 2.3: The number of nations with GDPs beginning with the digits one through nine. Data from the World Bank.

2.8 Examples

2.8.1 Benford's law and world GDP

In the late 19th and early 20th century, before computers were widely used, day-to-day calculation of common mathematical functions such as logarithms depended on using large tables of computed values. Several researchers, most notably Simon Newscomb (1881) and Frank Benford (1938) noticed that some pages of these books were more heavily used than others. Benford formulated what has become known as "Benford's law," stating that in certain data-sets (sometimes called scale-invariant data), the first digit of the observations tended to be distributed[6] $\log_{10}\left(\frac{d+1}{d}\right)$. For example, the data $\{123, 5810, 33, 900\}$ has the first digits $\{1, 5, 3, 9\}$. Benford argued that, for certain data,[7] these first digits should not be uniform, as one might expect, but according to this curious distribution.

The probability, under this distribution, a given observation will have d as its first digit is shown in table 2.8.1. In a data-set following Benford's law, more observations should start with a 1 than start with a 9, for example.

For a specific data-set, consider the gross domestic products of 183 nations (taken from Wikipedia's report of the World Bank data). The 183 nations reported, ranging from the United States with the highest GDP[8] to Kiribati with the smallest. The frequency of the first digits is shown in table 2.8.1.

First, enter these data into `R`.

```
> gdprod<-c(57,31,28,14,11,11,9,10,12)
```

[6] On the support $d \in \{1, \ldots, 9\}$ this is a proper probability mass function.
[7] The conditions are rather hard to pin down.
[8] The EU is calculated as separate nations, otherwise it would have a slightly higher GDP.

It would be possible to test if these data follow a uniform distribution, as we might otherwise guess. The discrete uniform distribution puts equal probability on each of the outcomes.

```
> unifprob<-rep(1/9,9)  # set up probabilities
> chisq.test(gdprod,p=unifprob)

        Chi-squared test for given probabilities

data:  gdprod
X-squared = 100.1311, df = 8, p-value < 2.2e-16
```

The very low p-value gives ample reason to reject the idea that these data are uniformly distributed. Alternatively, these data can be tested against the null hypothesis that they come from Benford's distribution.

```
> benprod<-1:9                                # setting up probabilities
> benprod<-log((benprod+1)/benprod,base=10)  # according to Benford's law
> chisq.test(gdprod,p=benprod)

        Chi-squared test for given probabilities

data:  gdprod
X-squared = 4.8807, df = 8, p-value = 0.7702
```

The p-value is very high, meaning there is no reason in these data to reject the null hypothesis that these data came from Benford's distribution.[9]

2.8.2 Archaeology

The analysis of animal bones from an archaeological site poses some problems.[10] In many cases, the assemblage is a jumble of bones; it is unclear how many individual animals an assemblage represents. Specialized measurements are employed to assess questions of faunal diversity or abundance. Two common methods are

NISP	Number of identified specimens (in a given taxonomic group)
MNI	Minimum number of individuals (in a given taxonomic group)

The number of identified specimens (NISP) is simply the count of all the bones which can be identified to a specific taxon. Minimum number of individuals is a count of the smallest number of animals from a given taxon which could account for an assemblage. The goal of this investigation is to determine if the two assemblages represent the same proportions of taxa. Both MNI and NISP are given for two species, Steller sea lion (*Eumetopias jubatus*) and the common seal (*Phoca vitulina*) were simulated and are shown below.

[9] This is not quite the same as proving it *did* come from Benford's distribution; it may be the sample size is insufficient to detect deviances from Benford's distribution, for instance.

[10] This example is based on material in Donald Grayson. (1984) *Quantitative Zooarchaeology.*

	Site 1	Site 2
Eumetopias jubatus		
NISP	266	419
MNI	20	22
Phoca vitulina		
NISP	516	558
MNI	24	27

These data are generated such that there are different numbers of animals at each site, but their ratio is similar (that is, the underlying population from which both samples are drawn have the same frequencies of animals). Thus, since these are not real data, we are assured that the null hypothesis is true.

First, reading in the data and assemble it into contingency tables.

```
> mm<-read.table(
      file='http://students.washington.edu/nesse/qerm514/data/marinemammal.txt',
      header=T)
> nisp.mat<-matrix(c(mm$ssl.nisp[1],mm$cs.nisp[1],mm$ssl.nisp[2],mm$cs.nisp[2]),2,2,byrow=T)
> mni.mat<-matrix(c(mm$ssl.mni[1],mm$cs.mni[1],mm$ssl.mni[2],mm$cs.mni[2]),2,2,byrow=T)
```

Now run the χ^2 test for the NISP counts:

```
> chisq.test(nisp.mat)

        Pearson's Chi-squared test with Yates' continuity correction

data:  nisp.mat
X-squared = 14.0048, df = 1, p-value = 0.0001823
```

And run the χ^2 test again for MNI counts:

```
> chisq.test(mni.mat)

        Pearson's Chi-squared test with Yates' continuity correction

data:  mni.mat
X-squared = 0.024, df = 1, p-value = 0.877
```

Two different measures, two different results! The first test, using NISP, shows highly significant differences between the first site and the second site. The second test, using MNI, finds no reason to reject the null hypothesis.

The problem here is the assumptions of the χ^2 model. Recall that the χ^2 test is built on the data coming from a multinomial distribution (here, since there are only two species, this is the binomial). That is, in this case, if there are N_j animals of either species at site j, the model employed by the χ^2 test is that the number of the Steller sea lions (*Eumetopias jubatus*) at site j is given by a binomial random variable with N_j and p_j (the null hypothesis is that $p_1 = p_2$).

Is that model true for NISP? If there are 10 animals of either species at a site, and the probability of being a Steller sea lion is 0.3, we might guess 3 animals would be sea lions, but would not be greatly surprised by 2 sea lions or 4 sea lions. It is like flipping a biased coin ten times (heads for

sea lion, tails for seal). It would unsurprising to get 2 heads or 4 heads if the probability of heads is 0.3.

If, on average, both species leave 100 identifiable bones, this would give a range of about 200-400 bones from 1,000 bones total. Are the bones also binomially distributed? No. We would be very surprised if, in flipping a coin 1,000 we got 200 heads one time, and repeating the experiment, got 400 heads. The data are simply not distributed according to a multinomial (or binomial) distribution.

It is a reasonable question then, is MNI really a multinomial event either? It turns out probably not, but it is a more reasonable approximation than NISP.[11]

The bottom line, just because the data are categorical counts does not mean the χ^2 test should be used. The real requirement is that the data be distributed according to a multinomial distribution (or the special case, binomial).

2.8.3 Superhero BMI

The body mass index (BMI) is a common measure healthy and unhealthy weight-height combinations in people. BMI is defined as the ratio of a person's mass M (in kilograms) to to the square of their height H (in meters).

$$BMI = \frac{M}{H^2} \tag{2.3}$$

For both men and women, the Centers for Disease Control define a BMI less than 18.5 is considered "underweight," from 18.5 to 25 is "normal," 25-30 is "overweight" and BMI over 30 is "obese." (Note these are statistical categories; individual variation may be higher than is reflected here.) These categories are the same for both men and women (although are different for children). The US population in 2002 had an average BMI (for both men and women) of 28.[12]

The comic-book producer Marvel, has on their website has a variety of statistics on their characters, including height and weight. So how does Marvel superheroes compare to the US population?[13] The height, weight and BMI of a variety of marvel characters is shown in the table below.

[11] I did a quick simulation, which seemed to indicate it tends to fail to reject more often than it should, whereas NISP tends to reject more often than it should. This is fairly consistent with what we might expect, since under the null NMI and NISP would both be in roughly the same proportions as the original counts (that is, the number of each species deposited). However NISP is bigger, in absolute terms, than the actual counts, while MNI is smaller. Thus NISP overemphasizes the difference, while MNI underemphasizes it.

[12] This number is from a CDC press release I found online. Other sources seem to have other numbers, probably reflecting the specifics of the population sampled.

[13] This question was inspired by work of Healy and Johnson, whose data is used here. Their work is available at `http://girl-wonder.org/papers/bmi.html`

Name	height(in)	height(cm)	wt(lbs)	wt(kg)	BMI	sex
Felicia Hardy	70	177.8	120	54.55	17.25	F
Destiny	67	170.18	110	50.00	17.26	F
Ms Marvel	71	180.34	124	56.36	17.33	F
Storm	71	180.34	127	57.73	17.75	F
Kitty Pryde	66	167.64	110	50.00	17.79	F
Mary Jane Watson-Parker	68	172.72	120	54.55	18.28	F
May Parker	65	165.1	110	50.00	18.34	F
Jean Grey	66	167.64	115	52.27	18.60	F
Jessica Drew	70	177.8	130	59.09	18.69	F
Janet Van Dyme	64	162.56	110	50.00	18.92	F
Yuriko Oyama	69	175.26	128	58.18	18.94	F
Angela Del Toro	68	172.72	125	56.82	19.05	F
Elektra Natchios	69	175.26	130	59.09	19.24	F
Sue Storm-Richards	66	167.64	120	54.55	19.41	F
Jessica Jones	67	170.18	124	56.36	19.46	F
Betty Brant	67	170.18	125	56.82	19.62	F
Misty Knight	69	175.26	136	61.82	20.13	F
Gwen Stacy	67	170.18	130	59.09	20.40	F
Natasha Romanova	67	170.18	131	59.55	20.56	F
Emma Frost	70	177.8	144	65.45	20.71	F
Scarlet Witch	67	170.18	132	60.00	20.72	F
Madame Hydra	69	175.26	140	63.64	20.72	F
Arachne	69	175.26	140	63.64	20.72	F
Yelena Belova	67	170.18	135	61.36	21.19	F
Psylocke	71	180.34	155	70.45	21.66	F
Bruce Banner	69	175.26	129	58.64	19.09	M
Bobby Drake	68	172.72	145	65.91	22.09	M
Cloak	69	175.26	155	70.45	22.94	M
Pete Wisdom	69	175.26	158	71.82	23.38	M
Gambit	73	185.42	179	81.36	23.67	M
Alex Summers	72	182.88	175	79.55	23.78	M
Pietro Maximoff	72	182.88	175	79.55	23.78	M
Reed Richards	73	185.42	180	81.82	23.80	M
Helmut Zemo	70	177.8	167	75.91	24.01	M
Peter Parker	70	177.8	167	75.91	24.01	M
Scott Summers	75	190.5	195	88.64	24.42	M
Johnny Storm	70	177.8	170	77.27	24.44	M
Magneto	74	187.96	190	86.36	24.45	M
Danny Rand	71	180.34	175	79.55	24.46	M
Sam Guthrie	71	180.34	180	81.82	25.16	M
Jonah J. Jameson	71	180.34	181	82.27	25.30	M
Ben Parker	69	175.26	175	79.55	25.90	M
Frank Castle	73	185.42	200	90.91	26.44	M
Matthew Murdock	72	182.88	200	90.91	27.18	M
Tchalla	72	182.88	200	90.91	27.18	M
Victor Von Doom	74	187.96	225	102.27	28.95	M
Piotr Rasputin	78	198.12	250	113.64	28.95	M
Nick Fury	73	185.42	221	100.45	29.22	M
Brian Braddock	78	198.12	257	116.82	29.76	M
Steve Rogers	74	187.96	240	109.09	30.88	M

There are several tests we can do on these data: First determine if the BMI for superheros is significantly different from the US population. Stated formally, this is a test if these are random draws from a normal distribution with mean 28.

First, read in the data.

```
> sh<-read.table(
```

```
     file='http://students.washington.edu/nesse/qerm514/data/superhero.txt',
     header=T)
```

This creates `sh` as a dataframe with the column headings as given in the `superhero.txt` file, namely `name, height_in, height_cm, wt_lbs, wt_kg, BMI, sex`. To test whether the superheros on average have a BMI of 28, it is equivelent to test if $BMI - 28 = 0$. Thus the t test,

```
> t.test(sh$BMI-28)

        One Sample t-test

data:  sh$BMI - 28
t = -10.8657, df = 49, p-value = 1.191e-14
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -6.73097 -4.62983
sample estimates:
mean of x
  -5.6804
```

The p-value here is very small, indicating with reasonable certainty the data did not come from a normal with mean 28. This output further indicates that the mean of the superhero BMI was less than 28.

One part of the first homework assignment will be a further analysis of these data.

2.8.4 Presidential height

Presidents are elected in the United States every four years, and considerable statistical effort is made to prognosticate the result. The usual approach made is to do extensive polling or other survey techniques, gathering opinions. Absent those data, however, a predictor might be made using other data. Here height of the candidates is compared to the outcome of the campaign.[14]

There is a natural pairing of the heights of the presidents and first runner-up in each election.[15] The data, presidential heights since 1888, are shown in table 2.4. These data were culled from Wikipedia.

First, read the data in.

```
> pres<-read.table(
      file="http://students.washington.edu/nesse/qerm514/data/president_height.txt",
      header=T)

> t.test(pres$winh,pres$runuph,paired=T)

        Paired t-test
```

[14] It is reasonable here to question the independence assumption of the t test: Since some presidents run more than once, it is reasonable to expect the data to be correlated. In a more formal analysis, it might be worth exploring this issue, but here we'll just ignore it.

[15] As a matter of history, until the 12th amendment (1804) was passed, the runner-up took the office of the Vice President. Although the election structure was different then (each member of the electoral college would get two votes rather than one), in 1796, the president and vice president were from different parties.

Year	Winner's name	Winner's height (m)	Runner-up's name	Runner-up's height (m)	Difference
2004	GWBush	1.8	Kerry	1.93	-0.13
2000	GWBush	1.8	Gore	1.84	-0.04
1996	Clinton	1.89	Dole	1.83	0.04
1992	Clinton	1.89	GHWBush	1.88	0.01
1988	GHWBush	1.88	Dukakis	1.67	0.21
1984	Regan	1.85	Mondale	1.8	0.05
1980	Regan	1.85	Carter	1.75	0.10
1976	Carter	1.75	Ford	1.85	-0.10
1972	Nixon	1.82	McGovern	1.85	-0.03
1968	Nixon	1.82	Humphery	1.80	0.02
1964	Johnson	1.92	Goldwater	1.83	0.09
1960	Kennedy	1.83	Nixon	1.82	0.01
1956	Eisenhower	1.79	Stevenson	1.78	0.01
1952	Eisenhower	1.79	Stevenson	1.78	0.01
1948	Truman	1.75	Dewey	1.73	0.02
1944	FRoosevelt	1.88	Dewey	1.73	0.15
1940	FRoosevelt	1.88	Willkie	1.85	0.03
1936	FRoosevelt	1.88	Landon	1.73	0.15
1932	FRoosevelt	1.88	Hoover	1.80	0.08
1928	Hoover	1.82	Smith	1.68	0.13
1924	Coolidge	1.78	Davis	1.83	-0.05
1920	Harding	1.83	Cox	1.68	0.15
1916	Wilson	1.80	Hughes	1.80	0
1912	Wilson	1.80	TRoosevelt	1.78	0.01
1908	Taft	1.82	Bryan	1.83	-0.01
1904	TRoosevelt	1.78	Parker	1.83	-0.05
1900	McKinley	1.70	Bryan	1.83	-0.13
1896	McKinley	1.70	Bryan	1.83	-0.13
1892	Cleveland	1.80	Harrison	1.68	0.12
1888	Harrison	1.68	Cleveland	1.80	-0.12

Table 2.4: Presidential heights and the height of the first runner-up since 1888. All measurements in meters.

```
data:  pres$winh and pres$runuph
t = 1.2642, df = 29, p-value = 0.2162
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01317996  0.05584662
sample estimates:
mean of the differences
             0.02133333
```

This is a t test with a null hypothesis that the difference is equal to zero. From these results, there is no reason to think height is a significant predictor of the winner of the presidential contest.

2.8.5 Growth randomization test

(Many thanks to Ian Taylor for sharing this example.) Randomizations tests can be used in a wide range of circumstances; most often the statistic being compared is the mean, however a wide range of statistics can be used. In this case, a parameter fit using a technique (nonlinear least squares, see 12) which has not yet been covered.

Dogfish (and many other fish[16]) grow at an ever-decreasing, modeled by the von Bertalanffy growth equation.

$$L_t = L_\infty(1 - \exp(-k(t - t_0))) \tag{2.4}$$

The equation relates length at age t, L_t, to three fit constants: k (a growth rate), L_∞ (asymptotic total length), and t_0 (accounting for length at birth). These constants for particular species are of great interest to fisheries managers for many reasons.

One noted dogfish researcher, Ian Taylor, estimated these parameters for two populations of dogfish: one in the 1940s and one modern times.[17] He shows that there has been a change in the asymptotic length L_∞, between the two time periods.

Parameters are estimated using nonlinear least squares, a technique which has not yet been presented. Later in this course several additional methods will be developed for comparison of models of this type. Nevertheless, it is worth asking the question: "is the difference between the two values of L_∞ found–one from the 1940s and one from 2000s–statistically significant?" Or, said another way, "what is the probability of getting such widely different L_∞ estimates by chance alone?"

The two made up data-sets, and their best-fit curves are presented in figure 2.3. These data, based solely on a visual examination, do call into question whether there really is a difference. (The commands to plot the data are shown further below.)

The two sets of fit parameters are shown below, along with the code used to generate it.

```
> ##Read in the data
> dogf<-read.table(file='http://students.washington.edu/nesse/qerm514/data/dogfish.txt',
```

[16] Fun fact: Dogfish are sharks, which sometimes are classified as fish and sometimes not. The term "fish," if it includes sharks, would not be a monophyletic group; however custom and tradition make sharks fish. So that question in Melville's Moby Dick about whether whales (mammals) are fish, it turns out, is not nearly as unreasonable as it initially sounds.

[17] His data and methods have not yet been published, so these data have been generated and techniques have been somewhat simplified.

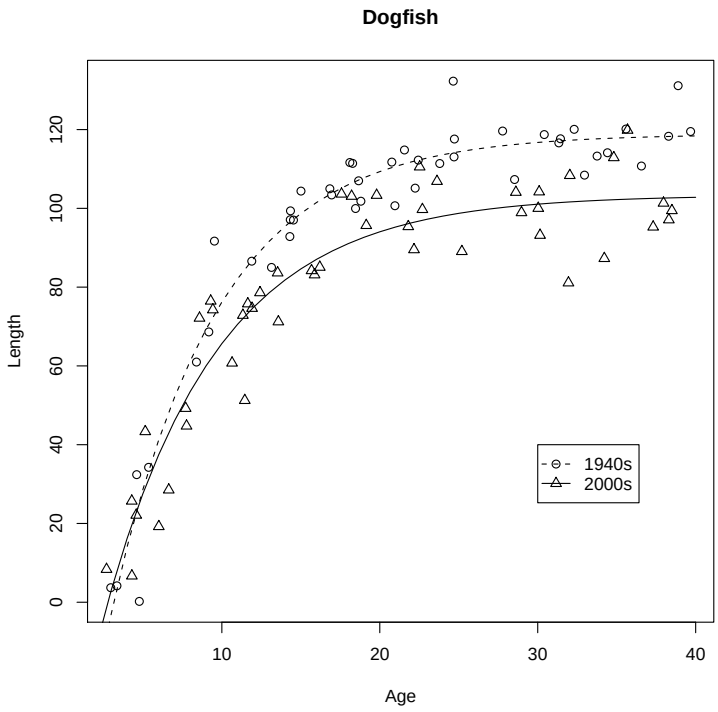


Figure 2.3: Generated data for Dogfish lengths and their best fit growth curves.

```
  header=T)
>
> ## make some variables which are easier to work with
> ag1<-dogf$age[dogf$group=='00s']
> ag2<-doff$age[dogf$group=='40s']
> Len1<-dogf$len[dogf$group=='00s']
> Len2<-doff$len[dogf$group=='40s']
>
> ##Fit the modern data
> fit1<-nls(Len1 ~ L*(1-exp(-k*(ag1 - t0))),start = list(L=120,k=0.15,t0=3))
> fit1
Nonlinear regression model
  model:  Len1 ~ L * (1 - exp(-k * (ag1 - t0)))
   data:  parent.frame()
       L        k       t0
103.3696   0.1397   2.7818
 residual sum-of-squares: 4456

Number of iterations to convergence: 4
Achieved convergence tolerance: 5.207e-06
>
> ##Fit the 1940s data
> fit2<-nls(Len2 ~ L*(1-exp(-k*(ag2 - t0))),start = list(L=120,k=0.15,t0=3))
> fit2
Nonlinear regression model
  model:  Len2 ~ L * (1 - exp(-k * (ag2 - t0)))
   data:  parent.frame()
       L        k       t0
118.8668   0.1498   3.1608
 residual sum-of-squares: 2336

Number of iterations to convergence: 2
Achieved convergence tolerance: 1.542e-06
>
> ## Plot the data
> plot(ag2,Len2,main="Dogfish",xlab="Age",ylab="Length")
> points(ag1,Len1,pch=2)
> legend(30,40,legend=c("1940s","2000s"),pch=c(1,2),lty=c(2,1))
> x<-1:40
> lines(x,predict(fit1,newdata=list(ag1=x)),lty=1)
> lines(x,predict(fit2,newdata=list(ag2=x)),lty=2)
```

The L_∞ values can be read here (called `L` in the code) as 103.37 cm for the modern data and 118.37 cm for the 1940s data, which is a difference of about 15 centimeters. Since the fitting method has not yet been covered, it is not important to be concerned about how it is fit, only that estimates of these parameters can be made from the data.

If there were no statistical difference between the two groups, it should not be unusual to get

the same difference or greater (in magnitude) of L_∞ fit values, if the data was drawn from the same distribution. Thus if the data were resampled, at random, into two new groups of the same size, it should not be surprising to get a difference of 15 or greater again.

Of course, one resampling does not really indicate how likely such a difference occurs. Instead, a large resampling (here 10,000 trials) is done, and the difference in L_∞ is recorded for each. This is implemented in `R` using a `for()` loop. Again, don't worry about the specifics of extracting coefficients or fitting; that will be covered in a later lecture.

```
## a variable to store the resampled differences in
resampled.differences<-NULL

## create vectors of all the ages together and all the lengths together
all.ages<-c(ag1,ag2)
all.lengths<-c(Len1,Len2)

## resample 10,000 times and calculate a difference in L_infty each time
for(zz in 1:10000){
    ## set up new samples
    first.group<-sample(1:length(all.ages),length(ag1))
    second.group<-sample(1:length(all.ages),length(ag2))
    ag1.resampled<-all.ages[first.group]
    ag2.resampled<-all.ages[second.group]
    Len1.resampled<-all.lengths[first.group]
    Len2.resampled<-all.lengths[second.group]

    ## fit both groups to the resampled data
    fit1.resampled<-nls(Len1.resampled ~ L*(1-exp(-k*(ag1.resampled - t0))),
                        start = list(L=120,k=0.15,t0=3))
    fit2.resampled<-nls(Len2.resampled ~ L*(1-exp(-k*(ag2.resampled - t0))),
                        start = list(L=120,k=0.15,t0=3))

    ## extract coefficients and calculate the difference in L_infty value
    L1<-as.numeric(coefficients(fit1.resampled)[1])
    L2<-as.numeric(coefficients(fit2.resampled)[1])
    resampled.differences[zz]<- L1 - L2
}
> length(resampled.differences[abs(resampled.differences)>=15])/10000
[1] 1e-04
```

The last line in this code indicates the approximate p value for this randomization test: 0.0001. Thus it is reasonable to conclude the values of L_∞ found in the 1940s and 2000s are indeed different.

Lecture 3

Linear Predictors

3.1 Main ideas

- Matrix notation for linear regression
- Solving least squares
- Properties of least squares
- Estimation of the variance
- Partition of sum of squares
- R^2, the coefficient of determination
- Making predictions based on new observations

3.2 Matrix notation

One of the most familiar statistical tools is ordinary linear regression via least squares. Traditionally, this model is written using β_0 and β_1 as fit parameters, x as a predictor observation, and y as the response observation. We observe pairs $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. The ordinary univariate least squares model relates x and y through the function

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{3.1}$$

The error, ϵ_i is modeled to be $N(0, \sigma^2)$ and independent of all other ϵ_j. Since Y_i is the sum of an observed x_i and some random noise ϵ_i, y_i is also a random variable. In fact, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. It is a crucial assumption of this model that all the ϵ_i terms have the same variance σ^2.

This model can be easily extended to multiple predictors. The data $\{(x_{1,1}, x_{1,2}, y_1), (x_{2,1}, x_{2,2}, y_2), \ldots, (x_{n,1}, x_{n,2}, y_n)\}$ can likewise be modeled by adding another fit parameter β_2. The resulting model, akin to equation (3.1), is

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i. \tag{3.2}$$

Likewise, if each observation is associated with k predictors, the response Y_i given the predictors x_i is modeled

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i. \tag{3.3}$$

Note that in the three equations above, all of three are really sets of n equations, since there is an equation for each i ranging from 1 to n. Said another way, there is an equation for each observation $(x_{i,1}, x_{i,2}, \ldots, x_{i,k}, y_i)$.

While this is not entirely unmanageable, matrix notation provides some tools for simplifying it. In a model with k predictors, first designate X the *model matrix*, also sometimes called the *design matrix*, as the $n \times p$ matrix with the predictors, where $p = k + 1$. That is

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n_2} & \cdots & x_{n,k} \end{pmatrix} \tag{3.4}$$

Further, put all p of the β_i into one vector β, and all of the Y_i into another vector Y.

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \qquad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \tag{3.5}$$

Using this notation, we can write all of the equations for ordinary linear regression as

$$Y = X\beta + \epsilon, \tag{3.6}$$

where $\epsilon \sim N(0, \sigma^2 I_n)$. This single equation is a bit easier to deal with than the collection of n univariate equations.

3.3 Solving least squares

One of the central goals of least squares is to fit the best vector β to equation 3.6. The word "best" is not really clear; in fact several methods exist which have different properties. However the most popular, for reasons which will be discussed later, involves minimizing the squared differences of the model's predictions $X\beta$ and the actual observations y. To visualize this, consider the case with a single predictor x.

As is shown in figure 3.1, the difference between the predicted value of the fitted linear model and the actual observations is known as the *residual*. The method for fitting the least-squares line is to find the values for the vector β which minimize the sum of the squared residuals. That is, find the values of β which minimize the sum of squared differences between the fit line and the observation. In general (for $p - 1$ predictors) this is shown in equation 3.7.

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \beta x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1}))^2 = (y - X\beta)'(y - X\beta) \tag{3.7}$$

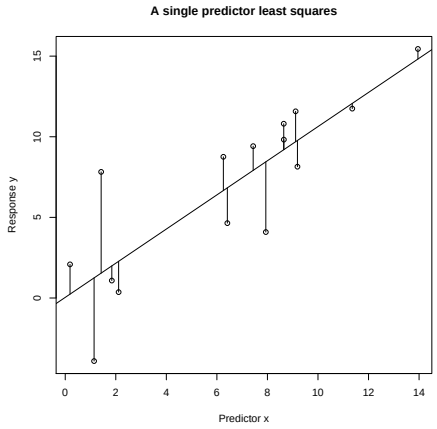


Figure 3.1: An ordinary least squares regression on data, with residuals to the line plotted. Data points are plotted as ∘, the best fit line is plotted going approximately through the data, while the residuals are shown vertically from the line to each point.

Note that since $(y - X\beta)$ is a column vector, $(y - X\beta)'(y - X\beta)$ is just the inner product of the vector with itself. One method of minimizing a function would be to systematically vary β, each time calculating the value of $(y - X\beta)'(y - X\beta)$. The following theorem, however, gives a computationally easier approach.

Theorem 3.3.1 (Least squares minimization)**.** *The function of beta $(y - X\beta)'(y - X\beta)$ is minimized when $\beta = (X'X)^{-1}X'y$, under the condition that the columns of X are linearly independent (meaning no predictor is exactly equal to another predictor or linear combination of other predictors)*[1]*.*

Proof. A necessary condition for a function to attain a minimum is that all of its partial derivatives are zero. Let $\hat{\beta}$ be the vector of β which minimizes equation 3.7. In matrix notation, this condition translates to

$$\frac{\partial}{\partial \hat{\beta}}(y - X\hat{\beta})'(y - X\hat{\beta}) = 0 \tag{3.8}$$

Note here that the differential operator $\frac{\partial}{\partial \hat{\beta}}$ operator will return a p dimensional vector, since it is taking a partial derivative with respect to each β_i. Expanding and differentiating the left side of

[1]This is a pretty reasonable assumption. You can imagine running into problems if, for instance, you entered a predictor twice in the matrix. Note that they don't have to be *statistically* independent, their correlation simply has to be neither 1 or -1.

equation 3.8 gives the following.

$$\begin{aligned}
\frac{\partial}{\partial\hat{\beta}}(y-X\hat{\beta})'(y-X\hat{\beta}) &= \frac{\partial}{\partial\hat{\beta}}(y'y-y'X\hat{\beta}-(X\hat{\beta})'y+(X\hat{\beta})'(X\hat{\beta})) && (3.9)\\
&= \frac{\partial}{\partial\hat{\beta}}(y'y-2\hat{\beta}'X'y+(X\hat{\beta}'(X\hat{\beta}))) && (3.10)\\
&= \frac{\partial}{\partial\hat{\beta}}(y'y-2\hat{\beta}'X'y+\hat{\beta}'(X'X)\hat{\beta}) && (3.11)\\
&= -2X'y+2X'X\hat{\beta} && (3.12)\\
& && (3.13)
\end{aligned}$$

By setting this equal to zero, it becomes clear that $X'y = X'X\hat{\beta}$. Since $X'X$ is assumed to be of full rank, based on the linear independence conditions of the the predictors, the matrix is invertible. Thus $\hat{\beta} = (X'X)^{-1}X'y$. This does not prove the theorem. Although the solution is unique, it remains to be shown that it is the maximum.

A critical point, identified as a location where the first derivative is zero, is a local minimum if the Hessian is positive definite. The Hessian matrix here, however, is (X'X), which is positive definite. Thus the point is known to be a minimum. □

3.3.1 $(X'X)^{-1}X'y$ as the MLE

Although the minimization of least squares was motivated by a specific statistical model, we never actually used the assumption $\epsilon \sim N_n(0,\sigma^2 I)$. In fact, in principle, least squares does give some kind of estimate for any choice of error function ϵ. However these estimates are not necessarily very good. For the ordinary least squares model, equation 3.6 however, the estimator $\hat{\beta} = (X'X)^{-1}X'y$ has several nice properties. First and foremost, it is the maximum likelihood estimator (MLE) of the true value β.

Theorem 3.3.2. *In the model $Y = X\beta+\epsilon$, where $\epsilon \sim N_n(0,\sigma^2 I)$, the maximum likelihood estimate is given by $\hat{\beta} = (X'X)^{-1}X'y$.*

Proof. Note that $Y \sim N_n(X\beta,\sigma^2 I)$ by the assumption of the model. Writing out the likelihood function for Y gives

$$L = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma^2}\right) \tag{3.14}$$

Taking the log of this expression does not change the value of β at which it is maximized. Taking the derivative of the log likelihood reduces to the same problem as minimizing the sum of squares.

$$\begin{aligned}
\frac{\partial}{\partial\beta}\log(L) &= \frac{\partial}{\partial\beta}\log\left(\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma^2}\right)\right) && (3.15)\\
&= \frac{\partial}{\partial\beta}\left(-\frac{n}{2}\log(2\pi)-n\log(\sigma)-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right) && (3.16)\\
&= -\frac{1}{2\sigma^2}\frac{\partial}{\partial\beta}(y-X\beta)'(y-X\beta) && (3.17)
\end{aligned}$$

This expression is equal to zero precisely when $\frac{\partial}{\partial\beta}(y-X\beta)'(y-X\beta) = 0$, which was the same problem faced in the minimization of the sum of squares. Further note that the Hessian $-\frac{1}{2\sigma^2}X'X$

is here negative definite, since it is a positive definite matrix times a negative constant. Thus the theorem is proved. □

3.3.2 Least squares as BLUE

The ordinary least squares regression is also the "Best Linear Unbiased Estimator," or BLUE, for the parameters β. This is true regardless of sample size. "Best" is perhaps a bit of an excessive superlative, since least squares is not really the right tool in every situation; in this context, "best" simply means that the least squares solution $(X'X)^{-1}X'y$ has the lowest variance of any unbiased estimator. This theorem is known as the Gauss-Markov theorem.

Theorem 3.3.3 (Gauss-Markov)**.** *Among the class of linear combinations of y, $c'y$, such that $Ec'y = \beta$, the OLS predictor $(X'X)^{-1}X'$ has the lowest variance as an estimator of any element of β.*

Proof. Note that to be an estimator of a single element of β it is necessary to eliminate the other elements of β. Let

$$s_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{3.18}$$

where the 1 is in the ith row of the vector. Thus, for a particular element β_i of β, let $c_{ols} = X(X'X)^{-1}s_i$.

Now consider a general function c of y such that $\mathrm{E}c'y = \beta_i$. Since c is constant, this can be written $c'\mathrm{E}y = c'X\beta = \beta_i$. From this, conclude that $c'X = s_i'$. Now write $c = c_{ols} + k$, for some vector k.

$$\begin{aligned} \mathrm{Var}(c'y) &= c'c\sigma^2 && (3.19)\\ &= (c_{ols}+k)'(c_{ols}+k) && (3.20)\\ &= c_{ols}'c_{ols} + k'k + 2c_{ols}'k && (3.21) \end{aligned}$$

However $2c_{ols}'k = 0$:

$$\begin{aligned} c_{ols}'k &= c_{ols}'(c - c_{ols}) && (3.22)\\ &= (X(X'X)^{-1}s_i)'(c - X(X'X)^{-1}s_i) && (3.23)\\ &= s_i'(X'X)^{-1}c - s_i'(X'X)^{-1}(X'X)(X'X)^{-1}s_i && (3.24)\\ &= s_i'(X'X)^{-1}X'c - s_i'(X'X)^{-1}s_i && (3.25)\\ &= s_i'(X'X)^{-1}X'c - s_i'(X'X)^{-1}X'c && (3.26)\\ &= 0 && (3.27) \end{aligned}$$

Therefore $\mathrm{Var}(c'y) = c_{ols}'c_{ols} + k'k$, which is clearly minimized when $k'k$ (which is always non-negative) is zero. □

3.3.3 Estimation of the variance σ^2

In addition to the parameters β, there is also one additional parameter in the model: the common variance σ^2. The MLE estimator for σ^2 is not difficult to derive (see below), however it has the unfortunate property that it gives a biased estimator (although the bias goes to zero for large sample size). Thus another, unbiased, estimator is derived.

Theorem 3.3.4. *The model $Y = X\beta + \epsilon$ where $\epsilon \sim N_n(0, \sigma^2 I)$, the MLE of σ^2 is $\frac{1}{n}(y-\hat{y})'(y-\hat{y})$.*

Proof. Proof follows the same pattern as theorem 3.3.2, with some different algebra. □

The expected value the MLE for σ^2 is not, however σ^2 (this is a bit of matrix algebra and probably requires Cochrane's theorem). Thus the estimator $\hat{\sigma}^2 = \frac{1}{n-p}(y-\hat{y})'(y-\hat{y})$ is used instead, where p is the number of parameters in the model.[2] Proving this result is a direct consequence of theorem 3.4.1, which has not yet been presented: Since the residual sum of squares is distributed $\sigma^2\chi^2_{n-p}$, it has expected value $\sigma^2(n-p)$.

3.4 Partition of the sum of squares

The estimate $\hat{\beta}$ can give rise to an estimated mean response $\hat{y} = X\hat{\beta}$, for the model matrix X. These $\hat{y}$ values are sometimes known as the "predicted" y, since they are approximating the expected value of each observation in y. The residual sum of squares, therefore, can be written as $(y-\hat{y})'(y-\hat{y})$. Another gage of variability of the data, which does not use the model, is the total sum of squares– the total variability around the mean of the data. Let $\bar{y} = 1\frac{1}{n}\sum y_i$ be a column vector of the mean of the elements of y. There is a useful partition of the total sum of squares into the residual sum of squares and the squared differences of the predicted values and the mean.

Theorem 3.4.1. *In the normal OLS model where $EY = X\beta$ and $\hat{y} = (X'X)^{-1}X'y$, the total sum of squares $(y-\bar{y})'(y-\bar{y})$ is equal to the sum of the model sum of squares $(\hat{y}-\bar{y})'(\hat{y}-\bar{y})$ and the residual sum of squares $(y-\hat{y})'(y-\hat{y})$. A corollary of this is that the residual sum of squares is independent of the model sum of squares.*

Proof. Note that when we solved the ordinary least squares problem, theorem 3.3.1, we minimized by setting the derivative of the residual sum of squares equal to zero. This resulted in the expression

$$X'(y-\hat{y}) = 0. \tag{3.28}$$

By premultiplying equation 3.28 by $((X'X)^{-1}X'y)'$, the result is still zero (although in 3.28 it was a zero vector, now it is a zero scalar).

$$\begin{aligned} X'(y-\hat{y}) &= 0 && (3.29)\\ ((X'X)^{-1}X'y)'X'(y-\hat{y}) &= 0 && (3.30)\\ (X(X'X)^{-1}X'y)(y-\hat{y}) &= 0 && (3.31)\\ \hat{y}'(y-\hat{y}) &= 0 && (3.32) \end{aligned}$$

Keep these two preliminary results in mind: $X'(y-\hat{y}) = 0$ and $\hat{y}(y-\hat{y}) = 0$.

[2] Beware many sources reserve the notation $\hat{\theta}$ only for MLE estimates of a parameter θ.

Now consider the total sum of squares $(y-\bar{y})'(y-\bar{y})$.

$$\begin{aligned}(y-\bar{y})'(y-\bar{y}) &= (\hat{y}-y-\bar{y}-\hat{y})'(\hat{y}-y-\bar{y}-\hat{y}) && (3.33)\\ &= (\hat{y}-\bar{y})'(\hat{y}-\bar{y})+(y-\hat{y})'(y-\hat{y})+2(\hat{y}-\bar{y})'(y-\hat{y}) && (3.34)\end{aligned}$$

Using equations 3.28 and 3.32, the last term, $2(\hat{y}-\bar{y})'(y-\hat{y})$, equals zero. Consider writing $(\hat{y}-\bar{y})'(y-\hat{y}) = \hat{y}'(y-\hat{y}) - \bar{y}'(y-\hat{y})$: the first of these terms is zero by equation 3.32. The second is a bit more subtle, however recalling that the first column of X is all 1's, a corollary of equation 3.28 is $1'(y-\hat{y})=0$. Since $\bar{y}(y-\hat{y}) = \left(\frac{1}{n}\sum y_i\right)1$, $\bar{y}'(y-\hat{y}) = \left(\frac{1}{n}\sum y_i\right)1'(y-\hat{y}) = 0$. This proves the result.

Applying Cochrane's theorem (threorem 0.3.6), this effectively shows the model variance and the residual variance are independent and χ^2 distributed. □

3.5 R^2, the coefficient of determination

A useful tool building on the partitioning of the sum of squares is the coefficient of determination, R^2.

Definition 3.5.1. *Coefficient of determination The coefficient of determination, R^2 is defined as*

$$R^2 = 1 - \frac{(y-\hat{y})'(y-\hat{y})}{(y-\bar{y})'(y-\bar{y})} = 1 - \frac{RSS}{TSS}, \tag{3.35}$$

where TSS is the total sum of squares and RSS is the residual sum of squares.

This definition should be fairly intuitive, since by the partitioning of the sum of squares,

$$(y-\bar{y})'(y-\bar{y}) = (\hat{y}-\bar{y})'(\hat{y}-\bar{y}) + (y-\hat{y})'(y-\hat{y}). \tag{3.36}$$

Dividing through by $(y-\bar{y})'(y-\bar{y})$ yields

$$1 = \frac{(\hat{y}-\bar{y})'(\hat{y}-\bar{y})}{(y-\bar{y})'(y-\bar{y})} + \frac{(y-\hat{y})'(y-\hat{y})}{(y-\bar{y})'(y-\bar{y})}. \tag{3.37}$$

The two terms on the right can be thought of as the proportion of the sum of squares taken up by the model and by the residual respectively. Thus, a bit of rearranging indicates that a reasonable interpretation of R^2 is the percentage of the total sum of squares "explained"[3] by the model.

Thus our intuition of R^2 should be that values close to 1 (note that R^2 as defined here can not exceed 1) indicate strong clustering about the regression line, while R^2 near 0 indicates wide variability about the line (see figure 3.5).

3.5.1 Adjusted R^2

This is a topic which is not going to be immediately relevant to this investigation, however will shortly make an appearance in `R`'s output. One criticism of R^2 is that models with more predictors necessarily have the same or greater R^2 even if those predictors are complete nonsense. Thus an adjustment is made to the R^2 for use in model selection, a later topic in this course.

[3] "Explained" is commonly used, however it is a bit misleading since it has connotations of causation.

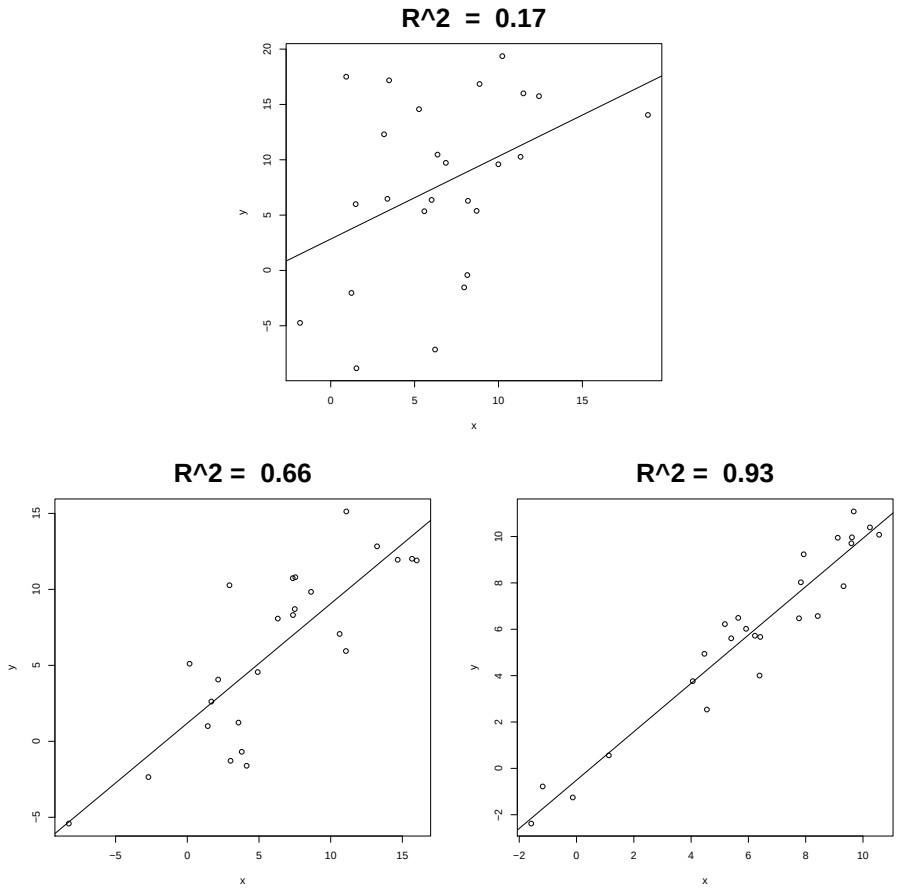


Figure 3.2: Plots of regressions with different R^2 values. Note that as R^2 gets closer to 1, the points more closely cluster around the line.

Definition 3.5.2. *Adjusted R^2 The adjusted R_a^2 is defined using the original R^2.*

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p} = 1 - \frac{Var(error)}{Var(total)} \tag{3.38}$$

This R_a^2 has, in a sense, a penalty for increasing the number of parameters in the model. Thus it is more easily adapted to model selection. While it still has some intuitive sense about the model fit, we not longer have the result that it is the percentage of the sum of squares taken up by the model.

3.6 Variations on a theme

3.6.1 Transformations of predictors

Often, for purely physical reasons, it is reasonable to suspect that a linear transformation of a variable should be substituted for the variable. Thus it is perfectly acceptable to enter x_i as one column of the design matrix, and the square (for example), x_i^2 as another column. In this case the regression solves the problem of fitting

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,1}^2 + \epsilon_i \tag{3.39}$$

While this equation is not linear in x_i, due to the square, the regression equation (which enters both x_i and x_i^2 as columns) still gives rise to a linear regression equation. One transformation, however, which is problematic is making one predictor out of linear combinations of other predictors, as this will result in a model matrix with less than full rank (and thus $(X'X)^{-1}$ would not exist).

3.6.2 Fixing the model through 0

It is possible to fix a regression line to go through zero; in matrix terminology, this is equivalent to dropping the first column of 1's from the matrix. However since the partition of the sum of squares is predicated on this column of ones existing, interpretation of R^2 (and the subsequently developed F test, later in this chapter) is problematic.

3.7 Making predictions of new observations

Once a linear model has been fit, it is sometimes useful to predict new observations based on new values of the predictors.[4] Conceptually, a new response y_{new} can be estimated given the new set of predictors $x_{1,new}$ through $x_{k,new}$, written as the vector x_{new}. The estimated response $\hat{y}_{new}$ is given by the multiplying the previously estimated parameters $\hat{\beta}$ by the new predictors. That is $\hat{y}_{new} = \hat{\beta} x_{new}$.

This should have an intuitive feel: once an estimate of β is made, the best guess for new responses based on new values of predictors is simply $\hat{\beta} x_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,new} + \cdots + \hat{\beta}_{p-1} x_{p-1,new}$. To predict a collection of new observations, the new values of the predictors can be collected into a new model matrix, X_{new}, so the vector of new predictions is $\hat{\beta} X_{new}$.

[4] This practice is not, however, without controversy. This issue will be explored further in the homework.

3.8 Examples

Name	section	description
Simple regression	3.8.1	Simplest possible regression, for reference on the commands.
Species area Biogeo.	3.8.2	Physical motivation for a transformation or predictors and response.
House prices	??	Multiple regressions, residuals and predictions.

3.8.1 A short example, with meaningless data

Suppose a researcher has response variable y and predictor variable x as given in the following table.

x	12.06	7.97	11.71	8.52	19.75	5.49	8.14	13.25	11.67	10.54
y	12.96	9.90	12.75	5.96	20.38	4.68	7.51	13.64	9.56	10.78

These data can be shown fit to the linear model $y = \beta_0 + \beta_1 x + \epsilon$ using the `lm()` command, and moreover, the resulting linear model object can be saved as a variable.

```
> simpleLinear.lm <- lm(y ~ x)
```

The linear object `simpleLinear.lm` can then be plotted or otherwise examined or manipulated. To get a summary of the linear model, the `R` command `summary` gives a great deal of information about the fit of the model, including estimates of the parameters β_0 and β_1.

```
> summary(simpleLinear.lm)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23002 -0.24689  0.06877  0.75980  2.31114

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1402     1.4260  -0.800    0.447
x             1.0958     0.1238   8.855 2.09e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.453 on 8 degrees of freedom
Multiple R-Squared: 0.9074,     Adjusted R-squared: 0.8959
F-statistic: 78.41 on 1 and 8 DF,  p-value: 2.087e-05
```

Later parts of these notes will go into more detail about the specifics of this output, however the estimates for β_0 and β_1 using ordinary least squares are shown in the `Coefficients` section under the heading `Estimate`. The fits here are the intercept, $\beta_0 = -1.1402$ and the slope $\beta_1 = 1.0958$.

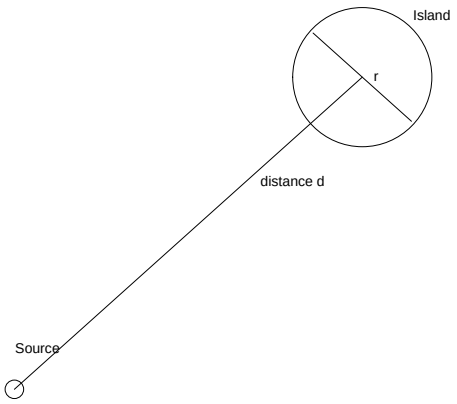


Figure 3.3: A source and sink model motivating the species-area relationship. The source location is sending new species out in all directions randomly, and those that arrive on the island are proportional to the arc intercepted.

3.8.2 Species-area relationship

One method of examining biodiversity is to simply count the number of species on an island within a given taxon. Observations on this have found that larger islands tend to have more species, while smaller islands have fewer. Proximity to the nearest continent, as one might guess, also seems to be a reasonable factor in predicting island species counts. One approach to modeling the relationship is to picture the island as a circle at some distance d from a source population (see figure 3.8.2).[5] The model assumes species come from a source population, which is at distance d from a sink island with area a. Imagine a circle centered at the source and of radius d. The percentage of species which arrive at the island are approximately proportional to the length of the arc of the circle which covers the island, divided by the total circumference of the circle. The diameter of the island is approximately equal to that arc, which in terms of area is $2\sqrt{a}/\pi$. The radius of the large circle is d, which makes the circumference $2\pi d$. Thus the probability of intercepting the island is approximately equal to

$$p \propto \frac{\sqrt{a}}{d}. \tag{3.40}$$

For a more complex island group, multiple distances might become important. Extinction rates likewise may be area dependent. However, one reasonable guess about the relationship between the

[5]More detailed discussions and work on this sort of question fall under the rubric of Biogeography. One of the early works on this topic was MacArthur and Wilson's *The Theory of Island Biogeography.*

number of species s within a specific taxon and the area z of an island is

$$s = ca^z. \tag{3.41}$$

where c is some fit constant (which may depend on distance) and z is another fit constant. Data on the number of species on an island and the island areas is certainly available. Logging both sides of the equation yields a linear equation which can then be fit with linear regression.

$$\log(s) = \log(c) + z\log(a) \tag{3.42}$$

The predictor here would then not be a but $\log(a)$, and the response would likewise be $\log(s)$. The data to test this prediction–a linear relationship between $\log(s)$ and $\log(a)$–can be found in the Kuril Islands, from a study done by the International Kuril Island Project.[6]

The Kuril Islands are a volcanic chain of small islands from Northern Japan to the Kamchatka Peninsula in Russia. Pietsch and colleagues surveyed, among other things, the number of terrestrial mollusk[7] species on 17 islands in the Kurils, as well as reported number of terrestrial mollusk species on several nearby landmasses. These data are an excellent opportunity to examine the species-area hypothesis outlined in equation 3.41 (see figure 3.8.2).

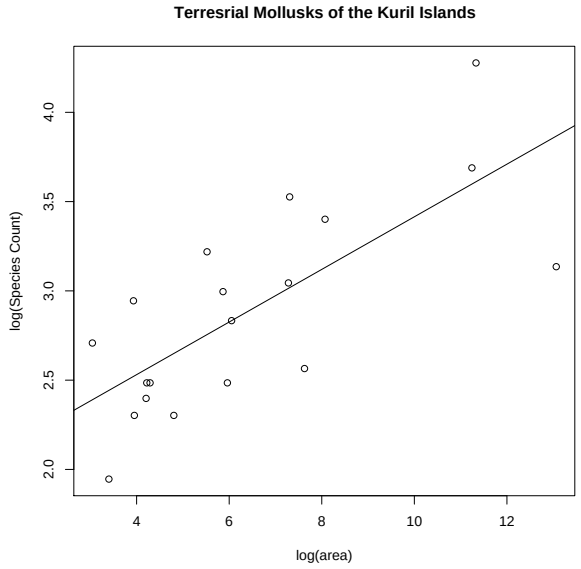


Figure 3.4: The number of species found on islands of specific areas in the Kuril Islands. Each point represents the log(area) of the island, and the log(number of terrestrial mollusk species) on the island)

[6]Pietsch et al. 2006. "Biodiversity and biogeography of the islands of the Kuril Archipelago", *Journal of Biogeography* 30.

[7]The phylum Mollusca has more than 112,000 members according to Wikipedia, of which 45 were identified in the Kurils

Although visually it appears to be a reasonable fit to the data, the next task is to determine the c and z constants of the model described in equation 3.41. To first plot the data, the plot shown in figure 3.8.2 can be generated with the following R commands:

```
> ## Import the data
> kurmol<-read.table(file="http://students.washington.edu/nesse/qerm514/data/kurilTerMol.txt",
    header=T)
> mols<-kurmol$ter.mol
> areas<-kurmol$area

> ## Generate a linear model of the data
> termol.lm<-lm(log(mols)~log(areas))
>
> ## Plot the data and linear model
> plot(log(areas),log(mols),main="Terrestrial Mollusks of the Kuril Islands",
>   xlab="log(area)",
>   ylab="log(mollusk count)")
> abline(termol.lm)
```

A summary of the fit can be generated using the `summary()` command.

```
> summary(termol.lm)

Call:
lm(formula = log(mols) ~ log(areas))

Residuals:
       Min         1Q     Median         3Q        Max
-7.311e-01 -2.780e-01  6.597e-06  2.943e-01  6.655e-01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.94146    0.22648   8.572 1.41e-07 ***
log(areas)   0.14735    0.03253   4.530 0.000296 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3962 on 17 degrees of freedom
Multiple R-Squared: 0.5469,     Adjusted R-squared: 0.5203
F-statistic: 20.52 on 1 and 17 DF,  p-value: 0.0002962
```

There is a lot of information presented here. Recall this is the summary for a model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. This is the same, of course, as fitting the model in equation 3.42, $\log(s) = \log(c) + z\log(a)$, by setting $\beta_0 = \log(c)$ and $z = \beta_1$. The coefficients β_0 and β_1 can be read off of the summary in the "Estimate" column of the Coefficients section. The estimate for the intercept, β_0 is 1.94146; by setting this equal to $\log(c)$, we get $c = \exp(1.94146) \approx 6.97$. The parameter z however, can be directly read off, $z = 0.14735$. Thus the model described in equation 3.41 is parameterized.

3.8.3 Housing prices

The value of houses is very much in the news these days.[8] Estimating the cost of a home relies on many factors, but two important components are floor space (area) and age. (The realtor's favorite, "location, location, location" is more difficult to quantify and is not included in these data.) A reasonable first approach to modeling the data is a linear model, with a response y of price, and predictors x_1 and x_2 of area in square feet and age in years (at the time of sale). These data are from Albuquerque, NM, in 1993.

Price	SQFT	Age	Price	SQFT	Age
2050	2650	13	1050	1680	13
2150	2664	6	1049	1900	34
2150	2921	3	934	1543	20
1999	2580	4	875	1173	6
1900	2580	4	805	1258	7
1800	2774	2	759	997	4
1560	1920	1	729	1007	19
1449	1710	1	710	1083	22
1375	1837	4	975	1500	7
1270	1880	8	939	1428	40
1250	2150	15	2100	2116	25
1235	1894	14	580	1051	15
1170	1928	18	1844	2250	40
1155	1767	16	699	1400	45
1110	1630	15	1160	1720	5
1139	1680	17	1109	1740	4
995	1500	15	1129	1700	6
900	1400	16	1050	1620	6
960	1573	17	1045	1630	6
1695	2931	28	1050	1920	8
1553	2200	28	1020	1606	5
1020	1478	53	1000	1535	7
1020	1713	30	1030	1540	6
850	1190	41	975	1739	13
720	1121	46	940	1305	5
749	1733	43	920	1415	7
2150	2848	4	945	1580	9
1350	2253	23	874	1236	3
1299	2743	25	872	1229	6
1250	2180	17	870	1273	4
1239	1706	14	869	1165	7
1125	1710	16	766	1200	7
1080	2200	26	739	970	4

The model being fit is therefore

$$y = \beta_0 + \beta_1 \underbrace{x_1}_{area} + \beta_2 \underbrace{x_2}_{age} + \epsilon \tag{3.43}$$

which of course can be written is matrix notation: $y = X\beta + \epsilon$. The first column of the X matrix is ones (since β_0 is not multiplied by anything), the next column is the $SQFT$ predictor, and the final column is the age predictor. The `model.matrix()` command, as shown below, displays the matrix X used in the linear regression.

Now to fit the model.

[8] Indeed, these housing prices made me feel nostalgia for the 90s.

```
> albHouses.lm<-lm(Price ~ SQFT + Age, data=hspr)
> summary(albHouses.lm)

Call:
lm(formula = Price ~ SQFT + Age, data = hspr)

Residuals:
    Min      1Q  Median      3Q     Max
-516.17  -87.75  -21.35   95.73  719.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.95644   86.57779   0.207   0.8364
SQFT         0.69364    0.04459  15.555   <2e-16 ***
Age         -4.21787    1.78472  -2.363   0.0212 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 184.1 on 63 degrees of freedom
Multiple R-Squared: 0.7992,     Adjusted R-squared: 0.7928
F-statistic: 125.3 on 2 and 63 DF,  p-value: < 2.2e-16
```

The `summary()` command for linear models will be examined more carefully in the next lecture. For this example, it is sufficient to examine the estimates for β. In particular, $\beta_0 = 17.96$, $\beta_1 = 0.69$ and $\beta_2 = -4.22$. To look more carefully at the fitting, first examine the model matrix.

```
> print(X<-model.matrix(albHouses.lm))  ##Note this could be done
   (Intercept) SQFT Age                 ##with two commands:
1            1 2650  13                 ##  X<- model.matrix(...)
2            1 2664   6                 ##  print(X)
3            1 2921   3
4            1 2580   4
5            1 2580   4
6            1 2774   2
7            1 1920   1
8            1 1710   1
9            1 1837   4
...                              ## Snip the output for brevity
62           1 1229   6
63           1 1273   4
64           1 1165   7
65           1 1200   7
66           1  970   4
attr(,"assign")
[1] 0 1 2
```

Just for demonstration purposes, it is possible to replicate the fitting routine for the linear model. From theorem 3.3.1, $\hat{\beta} = (X'X)^{-1}X'y$. In R this becomes

```
> solve(t(X)%*%X)%*%t(X)%*%hspr$Price
                 [,1]
(Intercept) 17.956437
SQFT         0.693641
Age         -4.217865
```

Note the use of the matrix multiplication command, transpose and inverse: `%*%`, `t()`, and `solve()`. This result is the same as is found in the `summary()` result.

In either case, the fit indicates, as expected, increasing areas indicates higher sale prices, and increasing age indicates lower prices. A reasonable question for a realtor to ask: did anyone pay "too much" for their house? A quick examination of the residuals, called by the `residuals()` command,[9] is an indicator. The residuals are plotted against their data point number (the order of the data) in figure 3.5

```
> house.resid<-residuals(albHouses.lm)
> plot(1:length(house.resid),house.resid,
       main="Residuals of house prices",
       xlab="House number",
       ylab="Residual")
```

[9]Many sources will also use the `resid()` command, which is identical for this model, but `residuals()` is a more general method.

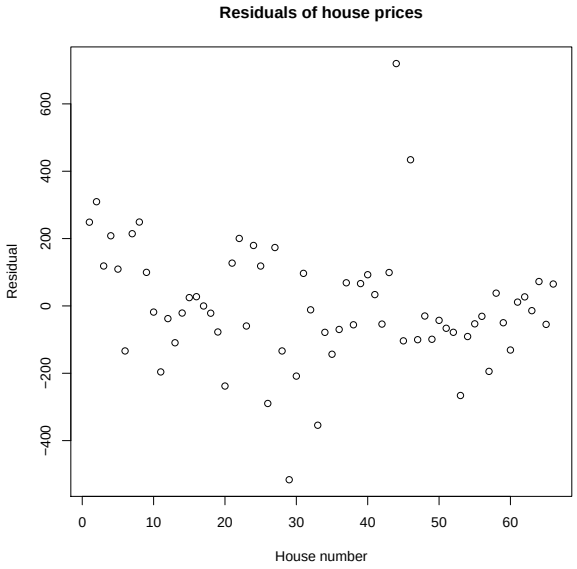


Figure 3.5: Residuals of the housing prices model, plotted against the order in which the data was given. Point 44 seems to indicate a higher payment than expected, while point 29 seems to represent an exceptionally low price for what is being sold.

Lecture 4

Inference on linear models

4.1 Main ideas

- t-tests
- F-tests
- Partial F tests
- β_i Confidence intervals
- Anova and Summary tables
- Confidence intervals on predictions

4.2 Background

Fitting the model generally means determining the set of $\hat{\beta}$ from the data. Interpreting the results (that is the $\hat{\beta}$ values) is still necessary. Intuitively, if for the simple model of $y = \beta_0 + \beta_1 x$, if x in reality has no influence on y, then the β_1 should be close zero. The natural question is 'what does 'close' mean?' More formally, what is the distribution of the β vector under the null hypothesis (although we still have to carefully formulate the null hypothesis)?

The estimate for β, written $\hat{\beta}$, is a vector of values, and using the usual ordinary least squares estimation, given by $\hat{\beta} = (X'X)^{-1}X'y$. Note that, prior to making an observation y, the response Y is a random variable. Thus $\hat{\beta}$ can be thought of as a random draw from a distribution given by $(X'X)^{-1}X'Y$ (note capital Y). Since Y is distributed multivariate normal, $(X'X)^{-1}X'Y$, as a linear transformation of Y, should also be multivariate normal (see theorem 0.3.3).

Theorem 4.2.1. *Distribution of $\hat{\beta}$ In the usual model, $Y = X\beta+\epsilon$, where Y is a vector of responses of length n and β is parameter vector of length k, the OLS estimate of β, $\hat{\beta} = (X'X)^{-1}X'Y \sim N_k(\beta, \sigma^2(X'X)^{-1})$.*

Proof. The proof follows directly from theorem 0.3.3. Since $Y \sim N_n(X\beta, \sigma^2 I_n)$, by theorem 0.3.3, $(X'X)^{-1}X'Y \sim N_k((X'X)^{-1}X'(X\beta), \sigma^2(X'X)^{-1}X'I_n((X'X)^{-1}X')')$. While unwieldy in this form, a bit of matrix algebra simplifies this to the result. (Verification of this is left as an exercise to the reader.)[1] □

Thinking about β as a random vector, it is possible to formulate several null hypotheses to test. We might test the null hypothesis that a given factor has no effect, after all the other elements have entered the model. We might ask if some collection of the β_is are all zero[2]. One particularly interesting subset of β_i to test is all except β_0, since this would be a test of the null hypothesis that *no* factor in the model has an effect.

Although several tests can be derived, the common approach to testing these is to use a version of the t-test for testing if an individual $\beta_i = 0$ and using an F test to check if all (except β_0) of the $\beta_i = 0$. A test if some of the β_i are zero is also done with an F test, but is usually called a "partial" F test.

4.3 t test

A simple null hypothesis for a single predictor x_a is that it has no influence on the response (once all the other factors have entered the model). Since the response is given by $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_a x_{ia} + \cdots + \beta_k x_{ik} + \epsilon_i$, this hypothesis could be equally well translated as the condition that $\beta_i = 0$.

Since, by theorem 4.2.1, the distribution of all the $\hat{\beta}_i$ is $N_p(\beta, (X'X)^{-1})$, where $p = k + 1$ is the number of parameters, the marginal distribution of $\hat{\beta}_i$ is given by

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2(X'X)^{-1}_{ii}), \tag{4.1}$$

where $(X'X)^{-1}_{ii}$ is the ith diagonal element of the $(X'X)^{-1}$ matrix. This can be rearranged to be

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2(X'X)^{-1}_{ii}}} \sim N(0, 1). \tag{4.2}$$

Since σ^2 is generally unknown, it is replaced by the unbiased estimator of it, $\hat{\sigma}^2 = \frac{1}{n-p}(y - \hat{y})'(y - \hat{y})$. The resulting expression is identical to the set-up for a t test.

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2(X'X)^{-1}_{ii}}} \sim t_{n-p} \tag{4.3}$$

Note that the degrees of freedom here come from the underlying model–n observations and k parameters to estimate. Under the null hypothesis $\beta_i = 0$, the relevant statistic is

$$t = \left| \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2(X'X)^{-1}_{ii}}} \right|. \tag{4.4}$$

[1] Don't you just hate lines like that?

[2] This is a different hypothesis than testing if each β_i is zero given the others separately; see example 4.7.2.

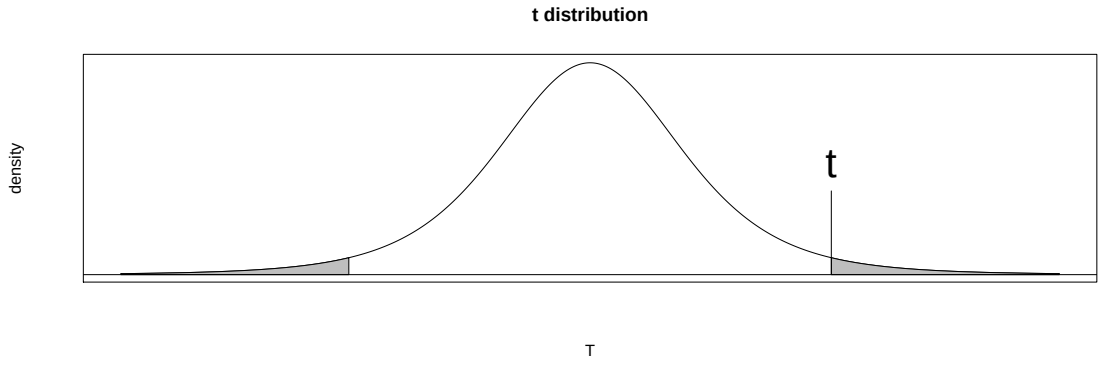


Figure 4.1: A t density, showing the value of $t = \left|\frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2(X'X)^{-1}_{ii}}}\right|$. The p-value of the test is the shaded area.

If this quantity is large (i.e. β_i is large relative to our estimate of its standard deviation, $\sqrt{\hat{\sigma}^2(X'X)^{-1}_{ii}}$, also known as the standard error), then the null hypothesis should be rejected. How large is too large, however, now has a specific meaning. Since this quantity is distributed t_{n-k}, the probability of a random draw from a t being as bad (that is, as far away from zero) or worse than the observed value (derived from equation 4.4) is the probability of being in the shaded region of figure 4.3.

4.3.1 Confidence interval construction

Using the same distribution for the $\hat{\beta}_i$ as the t test, equation 4.3, it is also possible to construct confidence intervals around the parameter estimate. A confidence interval is a common way of encapsulating the uncertainty in a parameter or value. The formal statement is a bit unwieldy: The $1-\alpha$ confidence interval is constructed according to rules which, if applied to random data under the null hypothesis, would construct intervals containing the true value $(1-\alpha)\times 100\%$ of the time. An incorrect, but common interpretation of a confidence interval is the "probability" that the true value lies in the interval.[3]

To construct the interval fix the value $t_{1-\alpha/2,n-p}$ such that if $T \sim t_{n-p}$, $P(|T| > t_{1-\alpha/2,n-p}) = \alpha$. See figure 4.3.1. This is, in a sense, the smallest value of t, from equation 4.4, which would be considered "significant" at the α level. However now the interest is constructing the interval around the estimate, not determining whether it equals zero.

Since any β_i such that $\left|\frac{\hat{\beta}_i-\beta_i}{\sqrt{\hat{\sigma}^2(X'X)^{-1}_{ii}}}\right| < t_{1-\alpha/2,n-k}$ would not be considered significant, the confidence interval can be constructed by rearranging this equation.

[3]Incorrrect at least in the frequentist interpretation of statistics. Under the frequentist philosophical viewpoint on statistics, the true value of a parameter is fixed. Once the confidence interval is found, the interval is likewise fixed. Ergo it does not make sense to talk aboout the probability one fixed number (the parameter) is between two other fixed numbers (the bounds of the confidence interval). An alternative philosophy of statistics, Bayesian statistics, holds that parameters can be thought of as random variables; however their construction of a confidence interval would be different as well.

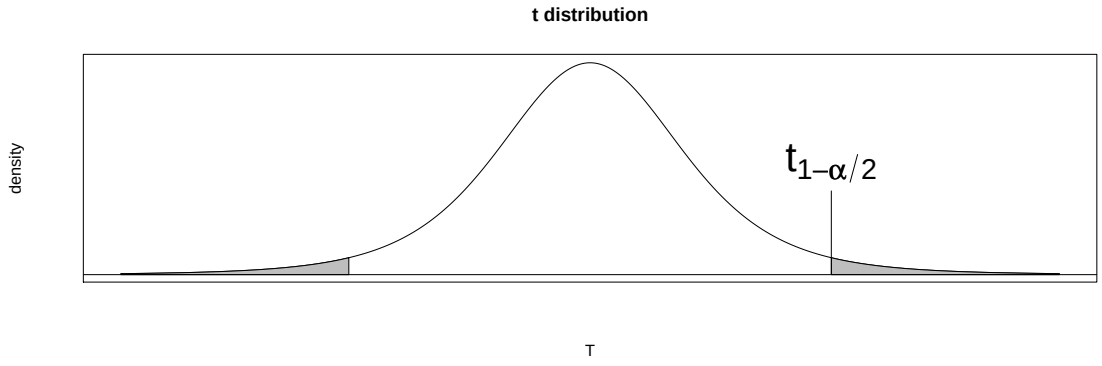


Figure 4.2: A t density, showing the point $t_{1-\alpha/2}$ such that $P(|T| > t_{1-\alpha/2,n-p}) = \alpha$.

$$\underbrace{\hat{\beta}_i - t_{1-\alpha/2}\sqrt{\sigma^2(X'X)^{-1}_{ii}}}_{\text{Lower CI bound}} \leq \beta \leq \underbrace{\hat{\beta}_i + t_{1-\alpha/2}\sqrt{\sigma^2(X'X)^{-1}_{ii}}}_{\text{Upper CI bound}} \tag{4.5}$$

4.4 F-tests

A generalization of the t test is the F test. It is useful for testing a wider range of null hypotheses on the same linear model $Y = X\beta + \epsilon$. The t test is limited to testing one parameter β_i at a time, while the F test can test the hypothesis that several β_i equal zero simultaneously. Purely by convention, "the" F test refers to the null hypothesis that all β equal zero except β_0 (the constant term). Other tests are referred to as "partial F tests."

The ordinary F test looks at the variance of the model predictions around the mean, divided by the residual variance. Partial F tests generalize this and look at the variance of a big model's predictions around a little model's fit again over the residual variance. Here a "big" model refers to a model with more predictors included, while a small model refers to a model with only a few predictors (or in the case of the ordinary F test, no predictors).

4.4.1 The main result

Consider the total sum of squares: $(y - \bar{y})'(y - \bar{y})$. In partitioning the sum of squares for the coefficient of determination, R^2, the total sum of squares could be seen as the sum of the model sum of squares plus the residual sum of squares. These, it turned out, were also independent. This result is actually sufficient to cover the ordinary F test, although a generalization is needed for the partial F test.

Theorem 4.4.1 (F-tests)**.** *Suppose $Y = X\beta + \epsilon$ is a model for Y (call it for this section "the large model") and $Y = X_r\beta_r + \epsilon$ be a reduced model such that the length of $0 <$ length$(\beta_r) <$ length(β). Suppose further that the columns of X_r are the identical to the first $r < p$ columns of X. This*

gives two sets of predictions: the $\hat{y}$ from the large model and $\hat{y}_r$ from the small model. The residual sum of squares of the small model $(y-\hat{y}_r)'(y-\hat{y}_r) = (y-\hat{y})'(y-\hat{y}) + (\hat{y}-\hat{y}_r)'(\hat{y}-\hat{y}_r)$, which are independent and each χ^2_{n-p} and χ^2_{p-r} respectively.

As such, under the assumption of the null model,

$$\frac{(SS_{small} - SS_{big})/(p-r)}{SS_{small}/(n-p)} \tag{4.6}$$

$$= = \frac{((y-\hat{y}_r)'(y-\hat{y}_r) - (y-\hat{y})'(y-\hat{y}))/(p-r)}{(y-\hat{y})'(y-\hat{y})/(n-p)} \tag{4.7}$$

$$= \frac{(\hat{y}-\hat{y}_r)'(\hat{y}-\hat{y}_r)/(p-r)}{(y-\hat{y})'(y-\hat{y})/(n-p)} \tag{4.8}$$

$$\sim F_{p-r,n-p} \tag{4.9}$$

As a side note, this can be put together with the partition of the sum of squares in theorem 3.28. The total sum of squares $(y-\bar{y})'(y-\bar{y})$ can be partitioned into sum of squares for each of the models. That is $(y-\bar{y})'(y-\bar{y}) = (y-\hat{y})'(y-\hat{y}) + (\hat{y}-\hat{y}_r)'(\hat{y}-\hat{y}_r) + (\hat{y}_r-\bar{y})'(\hat{y}_r-\bar{y})$. Moreover, these sums of squares are independent and have degrees of freedom $n-p$, $p-r$ and $r-1$ respectively.

Proof. As with the partition of the sum of squares in 3.28, note that finding the minimum sum of squares can be reduced to finding the $\hat{\beta}$ such that $X'(y-X\hat{\beta}) = X'(y-\hat{y}) = 0$. Note that this is actually k equations, all of which are set to zero. Thus $X_r'(y-\hat{y}) = 0$ is also true, since it is just a subset of those equations (but of course, the $\hat{y}$ estimates have not changed). Now premultiply this expression by the $((X_r'X_r)^{-1}X_r'y)'$.

$$0 = X_r'(y-\hat{y}) \tag{4.10}$$

$$= ((X_r'X_r)^{-1}X_r'y)'X_r'(y-\hat{y}) \tag{4.11}$$

$$= ((X_r(X_r'X_r)^{-1}X_r'y))'(y-\hat{y}) \tag{4.12}$$

$$= (X_r\hat{\beta}_r)'(y-\hat{y}) \tag{4.13}$$

$$= \hat{y}_r'(y-\hat{y}) \tag{4.14}$$

The rest of the proof is identical to theorem 3.28. As before $\hat{y}'(y-\hat{y}) = 0$. Consider the small model's residual sum of squares, $(y-\hat{y}_r)'(y-\hat{y}_r)$.

$$(y-\hat{y}_r)'(y-\hat{y}_r) = (y-\hat{y}+\hat{y}-\hat{y}_r)'(y-\hat{y}+\hat{y}-\hat{y}_r) \tag{4.15}$$

$$= (y-\hat{y})'(y-\hat{y}) + (\hat{y}-\hat{y}_r)'(\hat{y}-\hat{y}_r) - 2(\hat{y}-\hat{y}_r)'(y-\hat{y}) \tag{4.16}$$

$$= (y-\hat{y})'(y-\hat{y}) + (\hat{y}-\hat{y}_r)'(\hat{y}-\hat{y}_r) - 2\hat{y}'(y-\hat{y}) + 2\hat{y}_r'(y-\hat{y}) \tag{4.17}$$

$$= (y-\hat{y})'(y-\hat{y}) + (\hat{y}-\hat{y}_r)'(\hat{y}-\hat{y}_r) \tag{4.18}$$

The distribution and degrees of freedom are then found using Cochrane's theorem. To get this, first note that the rank of a hat matrix $X(X'X)^{-1}X'$ is the column dimension (p for X, and r for X_r). Thus $(y-\hat{y})'(y-\hat{y}) = y'(I_n - X(X'X)^{-1}X')y$ is a quadratic form of rank $n-k$, and $(\hat{y}-\hat{y}_r)'(\hat{y}-\hat{y}_r) = y'(X(X'X)^{-1}X' - X_r(X_r'X_r)^{-1}X_r')y$ is a quadratic form of rank $p-r$. □

4.4.2 The ordinary F test

Recall that the use of the F test is to compare nested models. The ordinary F test is a comparison of the model with all predictors to the model with no predictors (only a mean). The null hypothesis of the F test is that no predictor has an effect on the response; that is all responses are drawn from the same normal distribution (same mean, same variance).

Direct application of either theorem 4.4.1 or theorem 3.4.1 can be used to show that the difference between the total sum of squares (the sum of squared differences of observations around the mean of all observations, ignoring all the predictors) and the residual sum of squares (the sum of squared deviations of the observations around the model predictions) is χ^2_{p-1} and independent of the residual sum of squares, which is χ^2_{n-p}. See figure 4.4.2. Thus

$$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} = \frac{((y-\bar{y})'(y-\bar{y}) - (y-\hat{y})'(y-\hat{y}))/(p-1)}{(y-\hat{y})'(y-\hat{y})/(n-p)} \sim F_{p-1,n-p}. \quad (4.19)$$

There could be some intuitive sense about this. If TSS is very large, and RSS is small in comparison this would mean the responses vary wildly about the mean, but closely follow the model. This should be indicative of a good fit. Since a large TSS and small RSS would make equation 4.19 large, large values of the F statistic should be taken as evidence of good model fit.

Specifically, the p-value for an observation of F, is the probability that a $F_{p-1,n-p}$ random variable Q is greater than the observation.

$$p = P(Q \geq F) \quad (4.20)$$

In figure 4.4.2, the p-value is shown as the shaded area.

4.4.3 Partial F test

The partial F test[4] is a generalization of the ordinary F test. Instead of comparing the model with all predictors to one with no predictors (that is, just the mean), the full model is compared against a model with a fewer number of predictors. As before, appealing to theorem 4.4.1 shows that the sum of squares of the big model's predictions around the little model is independent of the sum of squares of the observations around the big model's predictions. Unlike the ordinary F test, the results are necessarily with several predictors; thus it is more difficult to visualize.

The basic statistic is very similar. The SS_{small} is the sum of squared differences from the small model predictions (comparable to the very small model of the mean in the ordinary F test). The SS_{large} is the sum of squared differences of observations from the larger model's predictions. Note that the smaller model must be nested within the larger model, ergo $\text{SS}_{\text{large}} < \text{SS}_{\text{small}}$, since by including more predictors, the sum of squares can not get larger.

The resulting F statistic, if the large model has p parameters and the smaller has r parameters both fit to n data, is

$$F = \frac{(\text{SS}_{\text{small}} - \text{SS}_{\text{large}})/(p-r)}{\text{SS}_{\text{large}}/(n-p)} = \frac{((y-\hat{y}_r)'(y-\hat{y}_r) - (y-\hat{y})'(y-\hat{y}))/(p-r)}{(y-\hat{y})'(y-\hat{y})/(n-p)} \sim F_{p-r,n-p} \quad (4.21)$$

[4]There is another test, the Chow test, which is occasionally called a partial F test at least according to Google. It is used to test if two regressions are identical.

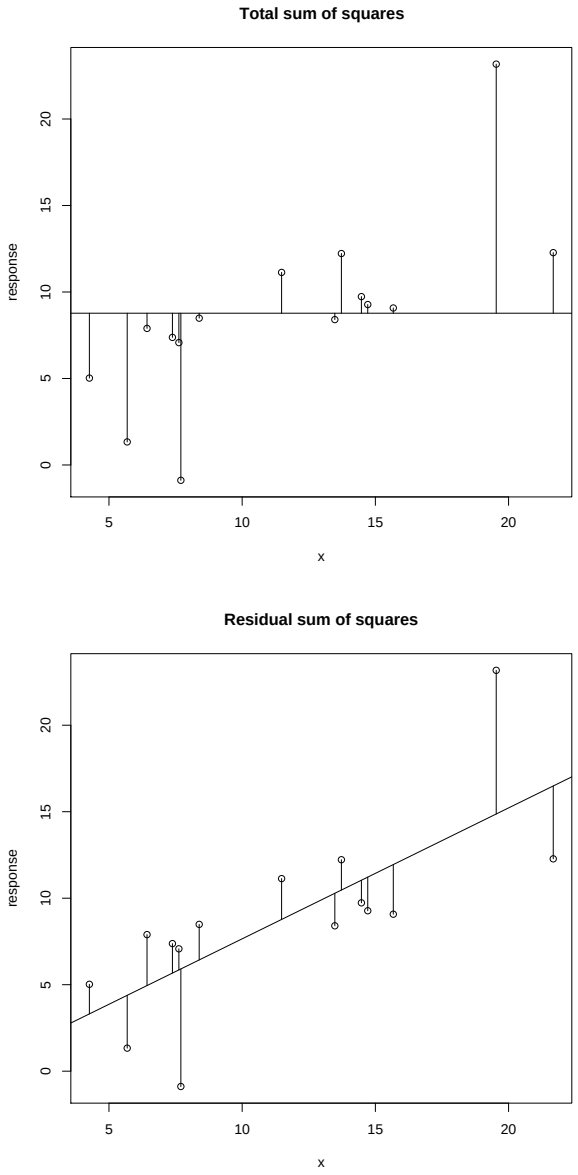


Figure 4.3: The total sum of squares and residual sum of squares are the squared differences from the mean (top) and model prediction (bottom) respectively. Visually, the sum of squares is the sum of the squared lengths of the vertical lines. These estimates, TSS and RSS, get combined to form the ordinary F test.

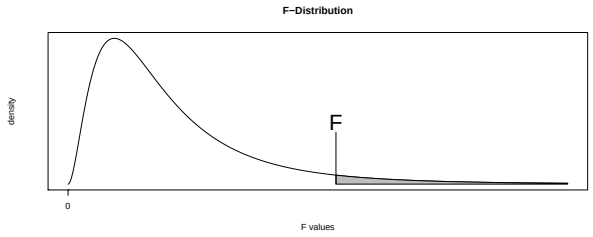


Figure 4.4: The p-value of the F test given an observation of F from the data is shown as the shaded area of this F distribution.

Note that this reduces to equation 4.19 when $r = 1$.

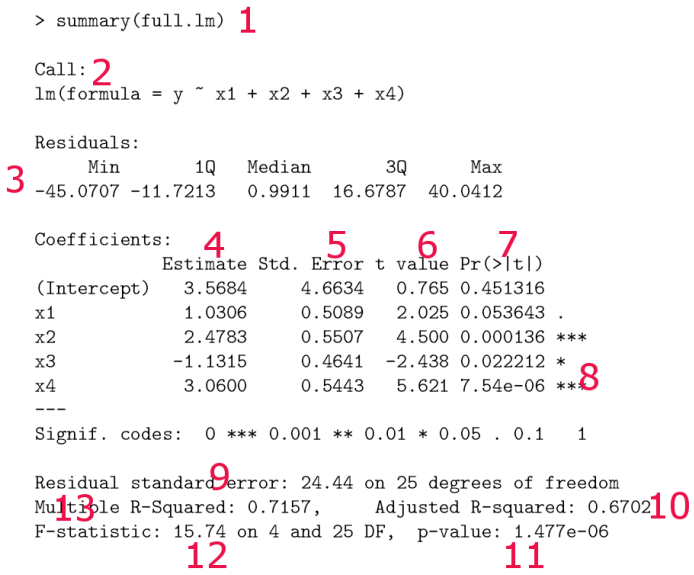
Equation 4.21 provides the necessary statistic to test the null hypothesis that the new predictors (that is, those in the large model but are not in the small) have no effect on the response. The intuition about this model is identical to that for the ordinary F test, and the p value is calculated in the same manner.

4.5 Reading anova and summary tables

Two of the commands in `R` which are commonly used to analyze a linear model are `summary()` and `anova()`. Both give a fair amount of information, which may be challenging to decipher.

4.5.1 The `summary()` command

Consider the parts of a summary table for a linear model:


> summary(full.lm)
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)
Residuals:
Min 1Q Median 3Q Max
-45.0707 -11.7213 0.9911 16.6787 40.0412
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.5684 4.6634 0.765 0.451316
x1 1.0306 0.5089 2.025 0.053643 .
x2 2.4783 0.5507 4.500 0.000136 ***
x3 -1.1315 0.4641 -2.438 0.022212 *
x4 3.0600 0.5443 5.621 7.54e-06 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Residual standard error: 24.44 on 25 degrees of freedom
Multiple R-Squared: 0.7157, Adjusted R-squared: 0.6702
F-statistic: 15.74 on 4 and 25 DF, p-value: 1.477e-06
1
2
3
4
5
6
7
8
9
10
11
12
13

1	Command	`summary`(linear model object) is what is entered in `R` to generate the table. The `summary()` command can actually summarize many other objects as well.
2	Call	Describes the linear model being summarized. Here it shows the regression formula is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$. The formula shown should be the model entered in the `lm()` function being summarized.
3	Residual description	For unweighted (equal variance) linear regression, this describes the minimum, quartiles, and maximum of the set of raw residuals $y - \hat{y}$.
4	Parameter estimates	The estimates of the parameters of the model, here β_0 through β_4. They are calculated as $(X'X)^{-1}X'y$.
5	Standard error	Standard error of the parameter estimate, calculated as $\sqrt{\hat{\sigma}^2(X'X)^{-1}_{ii}}$. To get the complete variance-covariance matrix, use `vcov()` (it can also be calculated using the `correlation = T` option of the `summary()` command).
6	t value	The t statistic, as shown in equation 4.3 with $\beta_i = 0$. Under the null hypothesis that x_i has no effect, this value should be near zero.
7	p-value	The two-sided t-test on the hypothesis that the parameter equals zero. Very low values indicate that it is highly unlikely that the parameter would be what it is by "chance" (that is, by the null hypothesis).
8	Significance	A quick visual guide to which p values are low (which may be useful if there is a long list of predictors). The guide to what the stars mean is given as `Signif. codes:` in the table. Three stars means less than 0.001, two stars is between 0.001 and 0.01, one star is between 0.01 and 0.05, a period means a p value between 0.05 and 0.1, and if the p-value is greater than 0.1 nothing is displayed.
9	Residual standard error	The unbiased estimate of the standard deviation of the error. It is calculated as the residual sum of squares over $n - p$.
10	Adjusted R^2	See section 3.5.1. It is a modification of the usual R^2 to account for number of parameters. It is calculated using the variance, rather than the usual R^2 method of sum of squares.
11	p-value	The p-value for the F test, testing the null hypothesis that no predictor in the model has any linear effect on the response.
12	F-statistic	The (ordinary) F statistic. Calculated using equation 4.19.
13	Multiple R^2	The coefficient of determination, indicating the proportion of the total sum of squares the model accounts for.

4.5.2 The `anova()` command

Consider the anova table output in `R` for the same linear model as the previous section:

```
> anova(full.lm)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 14481.9 14481.9 24.2534 4.528e-05 ***
x2         1  2949.2  2949.2  4.9392   0.03553 *
x3         1  1284.6  1284.6  2.1514   0.15491
x4         1 18869.3 18869.3 31.6011 7.537e-06 ***
Residuals 25 14927.7   597.1
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

1	Command	The `anova()` command can take any linear model (`lm()`) object or `aov()` object. It can also be used to compare two models which were not run at the same time, so long as they have the same response variable.
2	Degree of freedom	The degrees of freedom associated with each factor. Here, since the factors are all continuous predictors, each has one degree of freedom (one model parameter fit); in the section on anova, categorical predictors will be fit with multiple parameters, resulting in higher degrees of freedom. Note that the residuals have $n - p$ degrees of freedom (where n is the number of data and p is the number of parameters fit).
3	Sum of squares	The sum of squares of the difference between the model with the factors higher on the table to the model with factors higher and including the factor. That is, the `x3` factor sum of squares compares the model with `x2` and `x1` (and the mean) to the model with `x3`, `x2` and `x1` (and the mean).
4	Mean square	Each sum of squares divided by its degrees of freedom.
5	F value	The value of the F statistic for a partial F test, comparing the reduced model including higher factors to the model including higher factors and the factor on that line. (See sum of squares).
6	p-value	The probability of getting the described F value or higher, under the null hypothesis in which the response has no effect.
7	Significance	A quick visual guide to which p values are low (which may be useful if there is a long list of predictors). The guide to what the stars mean is given as `Signif. codes:` in the table. Three stars means less than 0.001, two stars is between 0.001 and 0.01, one star is between 0.01 and 0.05, a period means a p value between 0.05 and 0.1, and if the p-value is greater than 0.1 nothing is displayed.

4.6 Confidence intervals on predictions

Once a model is fit, it is possible to make predictions (based on the fit) of how the response variable would change according to new predictors (see section 3.7). These predictions, however, are not without uncertainty. The uncertainty arises from two sources–uncertainty in the estimates of β and uncertainty in any one observation.

From these two sources of uncertainty, two types of predictions come out. A confidence interval can be constructed around a single predicted observation or a different confidence interval can be constructed around the predicted mean observation. The confidence interval around the mean includes only variability due to the uncertainty in the estimates of β, while the confidence interval around a new observation also includes the observational variability around a line.

Theorem 4.6.1 (Prediction confidence intervals)**.** *Suppose in the model $Y = \beta X + \epsilon$, the least*

squares fit of $\hat{\beta} = (X'X)^{-1}X'y$ is used to predict a new observation $\hat{y}_{new}$ based on new values of the predictors x_{new}. Then the predicted mean observation $\hat{\bar{y}}_{new}$ will be distributed according to

$$\frac{\hat{\bar{y}}_{new} - \bar{y}_{new}}{\hat{\sigma}\sqrt{x'_{new}(X'X)^{-1}x_{new}}} \sim t_{n-p}. \tag{4.22}$$

This yields an α level confidence interval for the predicted mean response with x_{new} as predictors of

$$\underbrace{\hat{\beta}'x_{new} - \hat{\sigma}t_{1-\alpha/2}\sqrt{x'_{new}(X'X)^{-1}x_{new}}}_{Lower\ bound} \leq \bar{y}_{new} \leq \underbrace{\hat{\beta}'x_{new} + \hat{\sigma}t_{1-\alpha/2}\sqrt{x'_{new}(X'X)^{-1}x_{new}}}_{Upper\ bound}. \tag{4.23}$$

Likewise, the prediction for a single observation y_{new} is distributed

$$\frac{\hat{y}_{new} - y_{new}}{\hat{\sigma}\left(1 + \sqrt{x'_{new}(X'X)^{-1}x_{new}}\right)} \sim t_{n-p}. \tag{4.24}$$

The confidence intervals for a single new observation is given by

$$\underbrace{\hat{\beta}'x_{new} - \hat{\sigma}t_{1-\alpha/2}\left(1 + \sqrt{x'_{new}(X'X)^{-1}x_{new}}\right)}_{Lower\ bound} \leq \bar{y}_{new} \leq \underbrace{\hat{\beta}'x_{new} + \hat{\sigma}t_{1-\alpha/2}\left(1 + \sqrt{x'_{new}(X'X)^{-1}x_{new}}\right)}_{Upper\ bound}. \tag{4.25}$$

Proof. Note that $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$, thus $x'_{new}\hat{\beta} \sim N(x'_{new}\beta, \sigma^2 x'_{new}(X'X)^{-1}x_{new})$. Similarly $\hat{\beta}x_{new} + \epsilon \sim N(x'_{new}\beta, \sigma^2(1 + x'_{new}(X'X)^{-1}x_{new}))$. The result can then be obtained by the same reasoning as section 4.3.1. □

4.7 Examples

Name	section	description
Simple example	4.7.1	Simplest example using fake data
t and F test	4.7.2	Example of correlated predictors
Pisa	4.7.3	A simple yet complicated example

4.7.1 Simple example

There are quite a few ideas to go over. This example will go though all of them using 30 fake data (each with one response and three predictors), to give an idea of how to proceed with an analysis before tackling real data.

The data, shown below, can be fit with an multiple linear regression.

y	x1	x2	x3
118.59134	17.689361	19.4677955	44.497310
60.93338	7.286424	9.3739266	168.231765
133.51741	22.877860	42.2453026	20.619187
88.34969	8.695356	80.6232344	24.027538
128.41070	17.732708	62.7486259	15.605637
117.19637	14.233334	71.0855695	33.168792
104.83931	8.526150	69.4819795	141.901082
133.49469	17.751383	69.2936130	71.987626
94.44395	13.740090	0.6615558	28.237658
109.60687	11.973506	78.7123729	48.288079
75.73577	10.071377	7.1425624	51.937905
91.85394	11.461419	11.7036871	181.363809
148.06796	17.944899	85.0322473	1.135894
129.94741	14.668005	43.4940401	128.497851
148.82103	16.775813	99.0520237	28.766949
86.09425	14.435076	6.6290520	46.016558
71.36554	11.634327	9.3228299	24.522004
178.77304	25.572984	74.9031377	91.580912
115.70103	11.118161	94.2835786	46.865052
99.10088	16.344077	7.3222472	32.723340
99.37275	15.994169	9.9415650	30.327579
127.10369	16.422987	81.2280227	34.659225
72.38013	9.961278	10.6431198	67.251590
112.71650	15.326096	34.5524090	12.490110
106.09248	11.595912	51.5509824	9.094229
106.87356	13.805851	16.0222555	10.526501
128.91207	18.892054	28.1044331	21.109063
122.57865	21.929446	20.3101348	32.015707
88.44048	10.226970	81.6075200	9.022731
110.11387	18.281151	26.5844328	70.607060

First, read in the data, fit the model, and summarize it.

```
> fkdt<- read.table(
+     file="http://students.washington.edu/nesse/qerm514/data/fakeregressiondata.txt",
+     header=T)
> fk.lm<-lm(y~x1+x2+x3+x4,data=fkdt)
Error in eval(expr, envir, enclos) : object "x4" not found
> fk.lm<-lm(y~x1+x2+x3,data=fkdt)
> summary(fk.lm)

Call:
lm(formula = y ~ x1 + x2 + x3, data = fkdt)

Residuals:
    Min      1Q  Median      3Q     Max
-14.135  -6.445  -1.031   6.210  15.574

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.74930    6.66288   2.964  0.00642 **
x1           4.63344    0.36481  12.701 1.18e-12 ***
x2           0.42640    0.04845   8.801 2.83e-09 ***
```

```
x3           0.07127    0.03453   2.064  0.04915 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.31 on 26 degrees of freedom
Multiple R-Squared: 0.9072,     Adjusted R-squared: 0.8965
F-statistic: 84.73 on 3 and 26 DF,  p-value: 1.513e-13
```

There is a great deal of information already in the table. The estimates for the β parameters is in the table, as is the uncertainty associated with the errors. The t tests have each been performed on the parameters, indicating that each is significant (at some level). The null hypothesis for these tests is that each parameter β_i is zero (with the alternative $\beta_i \neq 0$), after all the other predictors have entered the model.

The estimate, standard deviation estimate (the standard error) can be combined to get confidence intervals around these parameters as well. Recall that confidence intervals require a α level, here taken to be the default value of 0.05.

```
> confint(fk.lm)
                   2.5 %     97.5 %
(Intercept) 6.0535499100 33.4450598
x1          3.8835627305  5.3833201
x2          0.3268074081  0.5259943
x3          0.0002873358  0.1422450
```

The `predict()` command is useful for predicting (surprise!) new data, or for determining the fitted values of the existing data.[5] In either case, confidence intervals of either new observation or mean can be made. First, develop confidence intervals around new predictions.

```
> set.seed(1)         #The next few lines are to make up some new predictors
> x1<-rnorm(6,15,4)   # to make predictions around.
> x2<-runif(6,0,100)
> x3<-rexp(6,1/50)
> new.data.madeup<-data.frame(x1,x2,x3)
>                                          ##now on to the predict command
> predict(fk.lm, new=new.data.madeup,interval="prediction")
       fit       lwr      upr
1 111.2584  93.61362 128.9032
2 111.0100  93.55734 128.4626
3 107.1137  88.97892 125.2484
4 144.9950 126.77824 163.2118
5 128.6726 111.06303 146.2822
6 120.7495 102.33136 139.1676
```

The command is almost the same to develop confidence intervals around the mean of the predictions.

[5] For finding the existing fitted values, often the `fitted.values()` command is used instead. There is no reason that I am aware of to prefer one over the other.

```
> predict(fk.lm, new=new.data.madeup,interval="confidence")
       fit      lwr      upr
1 111.2584 106.8341 115.6827
2 111.0100 107.4280 114.5920
3 107.1137 101.0223 113.2050
4 144.9950 138.6636 151.3265
5 128.6726 124.3909 132.9543
6 120.7495 113.8601 127.6388
```

Note that these confidence intervals are much smaller–they incorporate only the uncertainty in the estimate, rather than the uncertainty *and* in the observation.

Finally, for completeness, an anova is done below. Of course, given the summary table it is a forgone conclusion that the results of the partial F tests will all be significant.

```
> anova(fk.lm)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
x1         1 12165.8 12165.8 176.1797 4.345e-13 ***
x2         1  5092.9  5092.9  73.7525 4.562e-09 ***
x3         1   294.1   294.1   4.2595   0.04915 *
Residuals 26  1795.4    69.1
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

4.7.2 Example of t and F test

The t tests for individual parameters does not test the same thing as the F test (for all of the parameters). In some cases, the t test will give non-significant results for all of the parameters, while taken as a collection they are significant.

To illustrate this, the data generated here have a very high covariance. Thus, since the t test looks at the significance of the parameter after all of the other parameters have entered the model, the significance of the predictors is lost in the correlation. Said another way, the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is not a big improvement over either $y = \beta_0 + \beta_1 x_1$ or $y = \beta_0 + \beta_2 x_2$. However it would be a mistake to think that both β_1 and β_2 are zero.

The data generation process is shown here to demonstrate how correlated the data are. The command `set.seed(1)` is included to that this model will give the same results every time[6]. Note that the mean of `x2` is simply `x1`, and the variance of these is very small. As a result `x1` is highly correlated with `x2`.

```
> set.seed(1)
> x1<-rnorm(10,0,5)
```

[6]This has to do with how the random number generator works in R. The numbers generated are not really "random," but rather are generated by a complex function from a seed (the outcome of which, though deterministic, is hard to predict). Setting the seed starts this function at the same place, thus generates the same "random" numbers.

```
> x2<-rnorm(10,x1,.15)
> y<-rnorm(10,x1+x2,5)
```

Now running the linear model.

```
> co.lm<-lm(y~x1+x2)
> summary(co.lm)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7840 -1.0841  0.2643  1.3494  4.5128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6411     1.0961  -0.585    0.577
x1          -11.3021     7.1566  -1.579    0.158
x2           12.6054     7.2638   1.735    0.126

Residual standard error: 3.238 on 7 degrees of freedom
Multiple R-Squared: 0.7314,     Adjusted R-squared: 0.6547
F-statistic: 9.533 on 2 and 7 DF,  p-value: 0.01004
```

Note that the F test p-value is 0.01, while neither predictor, with p-values 0.16 and 0.13, is particularly significant. Highly correlated predictors is a common issue in data. This issue can be further elucidated via partial F tests.

```
> anova(co.lm)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x1         1 168.360 168.360 16.0540 0.005146 **
x2         1  31.583  31.583  3.0116 0.126255
Residuals  7  73.410  10.487
```

Here it is clearer that x_1 is useful as a predictor, however the benefit of x_2 is not clear. Of course, since in this ordering of the linear model (x_1 entering first, x_2 entering last), the p-value from the t test is identical for x_2 to the p-value from the last partial F test.

4.7.3 Leaning Tower of Pisa

Most examples in this course are doing things the way they should be done. This one is not, at least many people would argue that it is problematic. It was taken from another set of (online) regression notes,[7] and it is worth examining closely. The data are on the lean of the Leaning Tower of Pisa, as a function of time.

[7] I have no idea if the data are real or not.

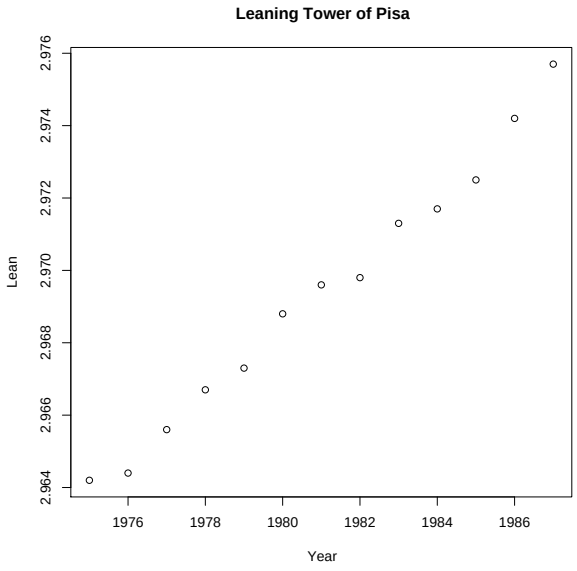


Figure 4.5: The lean of the Tower of Pisa as a function of time.

Year	Lean (m)
1975	2.9642
1976	2.9644
1977	2.9656
1978	2.9667
1979	2.9673
1980	2.9688
1981	2.9696
1982	2.9698
1983	2.9713
1984	2.9717
1985	2.9725
1986	2.9742
1987	2.9757

As figure 4.5 shows, the data do appear to be very linear. So plug and chug a linear model:

```
> pisa<-read.table(
      file="http://students.washington.edu/nesse/qerm514/data/pisa.txt",
      header=T)
> plot(pisa$year,pisa$lean,main="Leaning Tower of Pisa",
      xlab="Year",
      ylab="Lean")
> pisa.lm<-lm(lean~year,data=pisa)
> summary(pisa.lm)

Call:
```

```
lm(formula = lean ~ year, data = pisa)

Residuals:
       Min         1Q     Median         3Q        Max
-5.967e-04 -3.099e-04  6.703e-05  2.308e-04  7.396e-04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.123e+00  6.139e-02   18.30 1.39e-09 ***
year        9.319e-04  3.099e-05   30.07 6.50e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.0004181 on 11 degrees of freedom
Multiple R-Squared: 0.988,      Adjusted R-squared: 0.9869
F-statistic: 904.1 on 1 and 11 DF,  p-value: 6.503e-12
```

Unsurprisingly (after looking at the graph), the model is strongly significant. But is the model a good model for these data? Recall that one of the assumptions of linear regression was independence of responses given the predictor. That assumption may be highly questionable here.

Under the normal model, based on the independence assumption, it does not really matter how the predictors are arrayed.[8] Thus, if this were really a normal model, it should be no problem to remeasure the lean of the tower a few hundred times in one year–this should be adding information. Of course, intuitively, sampling every second for a minute adds no new information to this regression.

In fact, worse than that, it screws up the inference. Suppose a precocious young architecture student went nuts in 1978, and remeasured the lean of the tower every second for 30 seconds. Of course, the measurement would be the same each time (assuming the error arises from variability of the actual lean).

```
> #Set up a dataframe with 30 extra observations of 1978
> pisaExtra<-data.frame(c(pisa$year,rep(1978,30)),c(pisa$lean,rep(2.9667,30)))
> names(pisaExtra)<-names(pisa)
>
> ##run the regression again
> pisaExtra.lm<-lm(lean~year,data=pisaExtra)
> summary(pisaExtra.lm)

Call:
lm(formula = lean ~ year, data = pisaExtra)

Residuals:
       Min         1Q     Median         3Q        Max
-6.054e-04  2.638e-05  2.638e-05  2.638e-05  7.570e-04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

[8] This is not entirely true, as we'll see in chapter 9, but we'll let it go for now

```
(Intercept) 1.149e+00  2.707e-02   42.45   <2e-16 ***
year        9.188e-04  1.368e-05   67.16   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.0002221 on 41 degrees of freedom
Multiple R-Squared: 0.991,      Adjusted R-squared: 0.9908
F-statistic:  4511 on 1 and 41 DF,  p-value: < 2.2e-16
```

Note here that the regression coefficients changed only slightly (if, by chance, the point which had been repeatedly measured was unusual, these could have changed more dramatically). The significance of the regression went up enormously (several orders of magnitude), however.

What is going on here is a violation of the assumption of independence. Does this mean this model is useless? For a quick-and-dirty approach, no–it certainly tells the researcher the tower is leaning more and more, and gives an approximate trajectory for that lean. However fundamentally it is the wrong model and should be dealt with with a model which takes into account correlation, at least if proper inference is to be gained.

The take-home message of this example is that not all linear-looking plots should be used with linear regression[9] (except, perhaps, as a exploratory approach). This arises frequently in time series. Some time series are widely spaced enough so that the residuals are independent.[10] This is generally a physical assessment, however time series techniques such as looking at autocorrelation (correlation of the responses with the responses off-set by k years), can be employed. Whenever time arises, be very careful of the assumptions of the model being used–sometimes a time-series model is more appropriate.

[9]Never mind, however, that it is done all the time in a variety of contexts. In fact, the first proto-linear regression in the 17th century was done on data that had precisely this problem: trying to predict magnetic declination.

[10]In fact, some may argue that they are here, since they show no autocorrelation, but inarguably the data with many repeated observations is autocorrelated.

Lecture 5

Introduction to anova

5.1 Main ideas

- anova terminology and categorical predictors
- Design matrices and coding
- One way anova
- t-test interpretations and coding
- Brief look at contrasts, multiple comparisons

5.2 Categorical predictors

Categorical predictors are those which have a group, but not necessarily associated with a numerical value. Examples include species, class, sex, type, etc. One principle tool for handling factors is analysis of variance, or anova.

The analysis of variance (anova) is a classical tool in statistics, and principally refer to the use of categorical variables as predictors and a continuous response. The various sub-types of anova are shown in the table below.

One way Anova	In one way anova models, there is only one predictor, which is categorical.
Two-way anova	The addition of two or more categorical predictors allow for natural choices of models smaller than the full model.
Ancova	Analysis of covariance (ancova) is the mixture of continuous and categorical predictors.

Through cleverly encoding these variables (see coding section below), these models are all special cases of multiple regressions. That is, through some manipulations, it can be put into the same form as linear regression. Thus the same inferential tools–F tests and t tests–developed for continuous linear regression still hold true. In addition, however, there are a few tools and and interpretations which are specific to these anova models which require some attention.

5.3 Anova and multiple regressions

In the classical Analysis of Variance (anova) context all of the predictors are categorical and the response is a continuous normal random variable. Categorical predictors, recall, can be divided into groups (unlike, say a continuous predictor, which can be any number on a range). Let $i \in \{1, \ldots, n_j\}$ be the observation within groups $j \in \{1, \ldots, J\}$. Thus each response Y_{ij} is indexed by both group and observation.

Traditionally, the model for classical one way anova is written taking advantage of the group structure. Each Y_{ij} is normally distributed with mean μ_j (which is the same within a group but is allowed to vary between groups), and variance σ^2 (which is the same across all observations).

$$Y_{ij} = \mu_j + \epsilon_{ij} \tag{5.1}$$

The errors ϵ_{ij} are normal with mean zero and variance σ^2. In this context, the goal of the anova is to estimate each μ_j and (their common variance).

To more easily extend this into a multiple regressions context, μ_j is written as the sum of coefficients and predictors: $\mu_j = \beta_0 + \beta_1 x_1 + \cdots + \beta_{J-1} x_{J-1}$. Note that there are J coefficients, justas there are J means. By cleverly choosing predictors $x_1, \ldots, x_J$ finding the coefficients $\beta_0, \ldots, \beta_{J-1}$ is equivelent to finding the means $\mu_1, \ldots, \mu_J$.

The anova model can then be written in a familiar looking form.

$$Y_{ij} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{J-1} x_{J-1} + \epsilon_{ij} \tag{5.2}$$

"Cleverly" selecting the predictors, of course, is not something which should pass without comment. There are several methods for choosing predictors to be associated with each group. The underlying goal, however, is to be able to solve uniquely for each μ_j in terms of the coefficients $\beta_0, \ldots, \beta_{J-1}$. The process of choosing predictors is called "coding."

In most of regression, predictors are some kind of observation. Thus it may seem curious to say predictors can be chosen. To be clear, observations occur within a group and once the data are collected, those groups are fixed–they may not be reassigned. However it is reasonable to choose how each group is represented. That is, how each group is represented as values of $x_1, \ldots, x_J$ is a choice. The process of choosing how to represent groups as collections of predictors is called coding.[1]

A coding scheme is the choice of values for $x_1, \ldots, x_J$ which represent each group. One intuitive coding scheme is to represent the first group with $x_1 = x_2 = \cdots = x_J = 0$. Then represent the second group with $x_1 = 1$ and the other $x_i = 0$, the third group with $x_2 = 1$ and all other $x_i = 0$, and so on (see table 5.1).

Substituting these values of x into the anova model (equation 5.2), the means can be written in terms of various β_is. Since under this coding scheme, each β_i (for $i \neq 0$) represents the difference

[1] A great many things in mathematics and statistics are called coding. This has nothing to do with coding to maintain data integrity during noisy transmission, encryption, or the other concept also called coding.

Group	x_1	x_2	x_3	$\cdots$	x_{J-1}
1	0	0	0	...	0
2	1	0	0	...	0
3	0	1	0	...	0
$\vdots$			$\ddots$		
J	0	0	0	$\cdots$	1

Table 5.1: The treatment coding scheme, which is the default in `R` for coding unordered categorical predictors.

between the mean of group i and the first group (usually identified as the control of the experiment), this scheme is referred to as the "treatment" coding. Later, additional coding schemes will be considered, and a more careful treatment of the interpretation of the β_i values discussed.[2]

Just as in ordinary linear regression, this coding goes into a design matrix. Supposing there were two observations within each of three groups, the design matrix could be set up as shown below.

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \tag{5.3}$$

Notice that just as in the design matrix for ordinary continuous predictors (see equation 3.4), the first column of the matrix is all 1's. A difference, however, is that continuous variables each have one parameter associated with them; this is not true for categorical predictors. Here a single categorical predictor requires several parameters to enter the model.

Many other coding schemes are possible, and are discussed in a later section.

5.4 One way anova

Anova is built around categorical predictors and a continuous response. The simplest case of this is one way anova: a single categorical predictor giving rise to a continuous response.

The model for one way anova, equation 5.1, is fairly simple to describe. Each group has a different mean given by $\mu_j = \beta_0 + \beta_1 x_1 + \cdots + \beta_{J-1} x_{J-1}$. Observations $Y_{i,j}$ are the mean of the group plus some normal error. The estimators for the groups means are probably fairly intuitive.

$$\bar{Y_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{i,j} = \hat{\mu}_j \tag{5.4}$$

[2]For completeness, some books use the Kronecker product to write their coding. For us, table 5.1 is just a table indicating how a given categorical variable should be coded. Some authors, however, take the table to be a matrix, called the coding matrix, and take the model matrix to be the Kronecker product of the coding matrix and a vector of 1s (of length corresponding to the number of observations per group). This is silly for two reasons: first is, umm, why? It is unnecessary and gets you nothing (as far as I am aware). Secondarily, it does not generalize (at least as far as I know) to data with differing numbers of observations per level. We avoid it, as does Faraway, but for completeness mention it here.

The mean of a group in this model, is therefore estimated by the mean of the observations in that group. In fact, these are the maximum likelihood estimators (the proof of this is left to the reader).[3] For the estimate of the sample variance σ^2 which is shared by all groups, note that each group provides an independent estimate of the variance: $S_i = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2$. The mean of these individual estimates (weighted by their effective sample size, $n_j - 1$), therefore gives a combined estimate of the variance σ^2.

$$\frac{1}{N - J} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2$$

The methods of inference on the model parameters developed for ordinary linear regression are still useful here. In particular, by comparing the model $Y_{ij} = \mu_j + \epsilon_{i,j}$ (the ordinary one way anova model) with the simpler model $Y_{i,j} = \mu + \epsilon_{i,j}$ (note lack of the subscript on the μ, indicating every group has the same mean), an F test can be developed.

5.5 Inference

Since, at the fundamental level, the setup for linear regression and anova is the same, it should come as no surprise the same tools for inference are available. Individual parameters can be tested using a t test, as before, and groups of parameters can be tested with an F test. Since anova often assigns several parameters to a single predictor (unlike continuous regression which only assigns one), the interpretation and null hypothesis are a bit trickier here. Testing individual parameters will depend on the coding scheme used. Testing a single predictor, likewise, means testing multiple parameters, so an F test is a natural choice here.

5.5.1 Coding unordered factors

The `summary()` command outputs t-tests for each of the individual parameters (several of which may correspond to a single predictor). To understand what the summary table is describing, it is first necessary to consider the coding of the parameters in the model matrix.

Recall that in one way anova, a categorical predictor with J levels requires $J - 1$ parameters, together with the constant β_0. The default coding scheme for unordered categorical predictors in `R`[4] is the treatment contrast, `contr.treatment`. The treatment coding is described in table 5.1. The first group, gets zeros (except for the first column corresponding to β_0), the second gets 1's in the first place, and so on.

[3] For a double-plus good time, try proving it using matrix notation and an arbitrary coding scheme in the design matrix X.

[4] This is one of the few places where `R` differs from `S-PLUS`. The default coding for unordered categorical predictors in `S-PLUS` is the Helmert coding, described below.

Group	x_1	x_2	x_3	$\cdots$	x_{J-1}
1	1	0	0	...	0
2	0	1	0	...	0
3	0	0	1	...	0
$\vdots$			$\ddots$		
$J-1$	0	0	0	$\cdots$	1
J	-1	-1	-1	$\cdots$	-1

Table 5.2: The sum coding, `contr.sum`.

Consider the parameters in terms of the means of the group levels.

$$\begin{aligned}
\mu_1 &= \beta_0 \\
\mu_2 &= \beta_0 + \beta_1 \\
\mu_3 &= \beta_0 + \beta_2 \\
\vdots &\quad \vdots \\
\mu_J &= \beta_0 + \beta_{J-1}
\end{aligned}$$

Rearranging these equation to solve for the parameters results in

$$\begin{aligned}
\beta_0 &= \mu_1 \\
\beta_1 &= \mu_2 - \mu_1 \\
\beta_2 &= \mu_3 - \mu_1 \\
\vdots &\quad \vdots \\
\beta_{J-1} &= \mu_J - \mu_1
\end{aligned}$$

The interpretation of β_0 is the mean of the first group. The other parameters, β_1 to β_{J-1} is the difference between that mean and the first group. The first group is usually described as the control group. This coding scheme is designed for experiments in which multiple treatments are compared to a single control group.

A very similar coding scheme, SAS coding `contr.SAS`, uses the same setup but uses the last level of the group to be the control, rather than the first. It is named after the

Sum coding

The sum coding scheme, `contr.sum`, is also available in `R`. The coding, shown in table 5.5.1, can be written out as

$$\begin{aligned}
\mu_1 &= \beta_0 + \beta_1 \\
\mu_2 &= \beta_0 + \beta_2 \\
\vdots &\quad \vdots \\
\mu_{J-1} &= \beta_0 + \beta_{J-1} \\
\mu_J &= \beta_0 - \beta_1 - \beta_2 - \cdots - \beta_{J-1}.
\end{aligned}$$

This can be rearranged to solve for β_i.

$$\begin{aligned}
\beta_0 &= \frac{1}{n}\sum_{i=1}^{J}\mu_i \\
\beta_1 &= \mu_1 - \frac{1}{n}\sum_{i=1}^{J}\mu_i \\
\beta_2 &= \mu_2 - \frac{1}{n}\sum_{i=1}^{J}\mu_i \\
\vdots &\quad \vdots \\
\beta_{J-1} &= \mu_{J-1} - \frac{1}{n}\sum_{i=1}^{J}\mu_i
\end{aligned}$$

It is tempting, given this formulation, to say the parameter β_0 is the mean of all the data and β_1 through β_{J-1} are the differences of those levels from the grand mean. *This is true only when sample sizes are equal.* Since β_0 is really a mean of the group means, it is equal to the mean of the data only when the sample sizes are equal (otherwise levels with few samples are disproportionately represented).

Helmert coding

One last coding scheme in `R` is Helmert[5] coding. It is pretty awkward in interpretation, but in some cases may make the matrix computations easier.

5.5.2 Coding ordered factors

A special case comes up when levels of a group, while not numbered, are ordered. For example, in a drug trial, the dose might be reported as "low," "medium," and "high." While there are not numbers associated with them, there is a natural order. These levels are usually called "ordinal" since they have an order, although not necessarily specific numeric values (which would allow for linear regression).

[5]Named after the German geodesist Friedrich Robert Helmert.

Although there is no reason to suspect that the levels are equally spaced (that is, that if the predictors were continuous, the difference between adjacent levels would be equal to each other), it may make sense to at least look at that as a possibility or approximation. The polynomial coding, `contr.poly`, does just that.

The general formula for computing the values coded are a bit tricky. However they are set with a particular interpretation in mind. The β_0 parameter is the constant effect, β_1 is the linear effect assuming the levels are equally spaced, β_2 is the quadratic effect under the same assumption, β_3 is the cubic effect, and so on. Examples of linear and quadratic effect can be seen in figure 5.5.2.

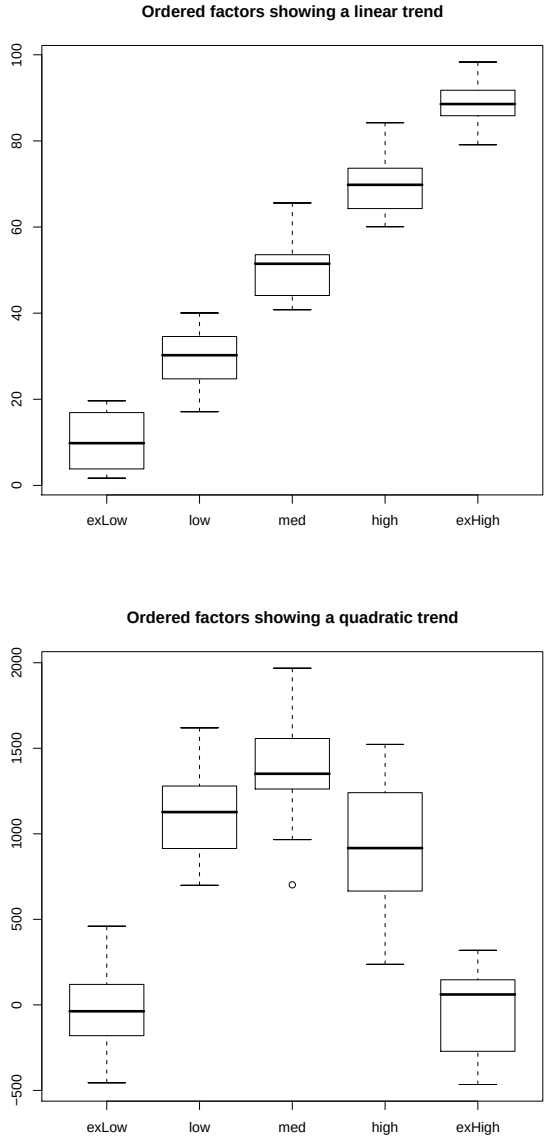


Figure 5.1: Plots showing linear and quadratic trend respectively in the responses to ordered predictors.

5.5.3 t tests and `summary()`

The `summary()` command in `R` performs t tests on individual parameters in the model. These tests no longer evaluate the null hypothesis of the predictor having no effect (if for no other reason than one predictor has multiple parameters associated with it). Instead, the t test evaluates a hypothesis which is dependent on the coding (see section 5.8.2 for an example). For the default coding, `contr.treatment`, the intercept (β_0) t test evaluates if the first (control) group has mean zero. The other t tests each evaluate whether that group's mean differs significantly from the control mean.

5.5.4 F tests and `anova()`

For one-way anova, there is only an ordinary F test since there is only one predictor (albeit one with multiple levels). The ordinary F tests examines the null hypothesis that the predictor has no impact on the response. Said another way, all the groups would have the same mean (i.e. $\mu_1 = \mu_2 = \cdots = \mu_J$).

5.6 Contrasts

Since the ordinary F test for anova (see section 5.5.4) is so simple, it is widely criticized as a poor way to do hypothesis testing. Casella and Berger call it "in many cases... silly, uninteresting, and not true" (pp 525). A common method to test more interesting hypotheses is to look at linear combinations of the group means. These linear combinations, subject to certain conditions, are called contrasts.

In fact, this lecture has already touched on contrasts: coding schemes result in specific sets of contrasts. The individual t tests of each parameter is, in reality, a test of a contrast.[6]

Contrasts of this sort can be used to test a wide range of hypotheses. Common approaches look at testing the differences between all pairs of levels (an extension of the treatment coding), or looking for the contrast with the most significant result.

5.6.1 Multiple comparisons

Something which is of great concern when performing many hypothesis tests is multiple comparisons. Perform enough tests, and something is bound to be significant. Thus several approaches have been developed for gaining inference. The Bonferroni correction is the crudest correction; if m tests are performed, set the significance for any one test to be α/m. This tends to be very conservative, since the tests of contrasts are not generally independent, thus this level of significance is too low (and what would otherwise be significant results are missed). Improved tests have been developed for specific types of tests, most notably Tukey's Honest Significant Difference (Tukey's HSD) for comparing all pairs of levels, and Scheffé's method for larger sets of comparisons.

[6] Well, most of them are contrasts; there are some technical requirements which have to be met to be called a contrast, which need not be met in a coding scheme.

5.7 Useful R commands

	Description
`summary()`	As before, summarizes the linear model
`anova()`	Outputs the anova table for the linear model
`model.matrix()`	Outputs the model matrix for the linear model
`C(...,contr=...)`	Sets the coding for a predictor
`factor()`	Codes as a factor, rather than another type of predictor
`as.factor()`	Coerces the argument to be read as a factor
`ordered()`	Sets up an ordered factor
`as.ordered()`	Coerces the argument to be an ordered factor.

5.8 Examples

Example	Section	Purpose
Made up data	5.8.1	Straight-forward one-way anova
Archaeological metals	5.8.2	A simple one-way anova, multiple comparisons.
Spices	5.8.3	An example of a nonsignificant anova.

5.8.1 Coding

The data here are fabricated to show the commands and interpretation to run an anova. Suppose a drug company wants to test the effectiveness of three new drugs, B, C, and D, against the standard treatment, drug A. The company contracts with the local hospital to treat statisticsitis[7] patients with one of the drugs, chosen at random. The time to recovery is measured, and displayed in the table below.

[7] Literally: Inflammation of the statistics

Days to recovery	Drug	Days to recovery	Drug
8.8106192	A	4.0493971	C
7.3623061	A	0.2740940	C
10.4241730	A	1.3844366	C
5.6550677	A	2.8515698	C
6.6850115	A	3.3928603	C
7.6600876	A	3.9697618	C
5.3163836	A	2.5648633	C
1.1929159	A	4.0753929	C
11.4458694	A	4.6949911	C
7.2083609	A	6.5454019	C
6.5708941	A	6.2367588	C
5.4852463	A	0.6774945	C
9.8982939	A	4.2543286	C
6.9376078	A	0.2261682	C
6.3433171	A	3.1580758	C
3.8387584	B	5.7889486	D
2.9465462	B	9.9897825	D
3.8132138	B	9.6619585	D
2.3438506	B	11.1091502	D
2.9738328	B	7.8405725	D
5.3572439	B	12.4711434	D
5.4088638	B	9.4994639	D
0.9528648	B	9.4628488	D
5.6994761	B	7.3552613	D
0.7235955	B	9.7169175	D
4.3220921	B	11.0553577	D
4.8864993	B	5.7463943	D
8.2860956	B	5.8609574	D
0.3906188	B	9.2991228	D
1.5055242	B	8.2535578	D

These means of each treatment group may, by visual examination (see figure 5.2) differ from each other. This visual determination is borne out by the anova table.

```
> st<-read.table(
      file="http://students.washington.edu/nesse/qerm514/data/statitisanova1.txt",
      header=T)
> boxplot(st$res~st$drug,
      main="Drug treatments",
      xlab="Drug",
      ylab="Days to recovery")
> st.lm<-lm(res~drug,data=st)
> anova(st.lm)
Analysis of Variance Table

Response: res
          Df Sum Sq Mean Sq F value    Pr(>F)
drug       3 342.39  114.13  24.207 3.563e-10 ***
Residuals 56 264.03    4.71
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The appropriate contrast for this model is probably the default coding, `contr.treatment`, in which the first parameter represents the mean of the first group, while the remainder show the

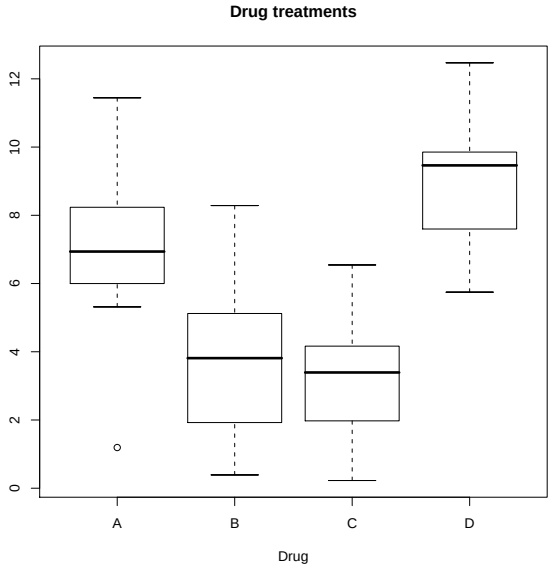


Figure 5.2: A boxplot of drug treatments for the example in section 5.8.1.

difference between each group and the first. Thus it is appropriate to use t tests to determine if each mean differs from the first (standard treatment).

```
> summary(st.lm)

Call:
lm(formula = res ~ drug, data = st)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9402 -1.4882  0.1992  1.1676  4.7228

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1331     0.5606  12.723  < 2e-16 ***
drugB        -3.5698     0.7929  -4.502 3.45e-05 ***
drugC        -3.9094     0.7929  -4.931 7.67e-06 ***
drugD         1.7410     0.7929   2.196   0.0323 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.171 on 56 degrees of freedom
Multiple R-Squared: 0.5646,     Adjusted R-squared: 0.5413
```

```
F-statistic: 24.21 on 3 and 56 DF,  p-value: 3.563e-10
```

These results can be interpreted, contingent on the coding. The `(Intercept)` represents the mean of the first group, which is unsurprisingly, not zero. The `drugB` coefficient estimate of -3.57 indicates that Drug B cures patients, on average, 3.57 days more rapidly than the standard treatment. The t test associated with Drug B ($p = 3.45 \times 10^{-5}$) strongly suggests this difference is significant. Likewise, Drug C (according to its coefficient) outperforms the standard treatment by almost four days. Drug D seems to do significantly worse than the standard treatment, as evidenced by the t test associated with the `drugD` coefficient.

Note that this summary table does *not* give the significance of differences between all drugs, only between each drug and the first (standard treatment).

Since the sample sizes are equal here, it might also be reasonable to try sum coding. Use the capital `C()` command to change coding: note that the lm command includes `C(drug, contr=contr.sum)` instead of just `drug`.

```
> st.lm<-lm(res~C(drug,contr=contr.sum),data=st)
> summary(st.lm)

Call:
lm(formula = res ~ C(drug, contr = contr.sum), data = st)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9402 -1.4882  0.1992  1.1676  4.7228

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5.6985     0.2803  20.329  < 2e-16 ***
C(drug, contr = contr.sum)1   1.4345     0.4855   2.955  0.00457 **
C(drug, contr = contr.sum)2  -2.1353     0.4855  -4.398 4.94e-05 ***
C(drug, contr = contr.sum)3  -2.4748     0.4855  -5.097 4.22e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.171 on 56 degrees of freedom
Multiple R-Squared: 0.5646,     Adjusted R-squared: 0.5413
F-statistic: 24.21 on 3 and 56 DF,  p-value: 3.563e-10
```

5.8.2 Archaeological metals

Archaeological investigations often work to identify similarities and differences between sites. Artifacts from one site may be compared to another using material differences between them, identifying one form or cultural entity from another. Analysis of the materials used in the making of pots, for instance, can be used to identify different manufacturing techniques (and thus differentiating between artifacts).

The percentage of iron found in pottery from four Roman-era sites in Britain is shown below. The four sites are Llanederyn (L), Caldicot (C), Island Thorns (I), and Ashley Rails (A).

Fe	Site
7.00	L
7.08	L
7.09	L
6.37	L
7.06	L
6.26	L
4.26	L
5.78	L
5.49	L
6.92	L
6.13	L
6.64	L
6.69	L
6.44	L
5.44	C
5.39	C
1.28	I
2.39	I
1.50	I
1.88	I
1.51	I
1.12	A
1.14	A
0.92	A
2.74	A
1.64	A

These data can be plotted using a boxplot (figure 5.3). The ultimate goal of the anova is to determine if the sites differ significantly in the metal they use.

```
met<-read.table(
    file="http://students.washington.edu/nesse/qerm514/data/metals.txt",
    header=T)
boxplot(met$Fe~met$Site,
    main="Percent Iron form pottery in 4 archaeological sites",
    xlab="Site",
    ylab="Percent Fe")
```

Note that there are different numbers of pottery sherds[8] tested at each site. A simple anova will suffice to test the hypothesis that each of the sites has the same mean.

```
> met.lm<-lm(Fe~Site,data=met)
> anova(met.lm)
Analysis of Variance Table

Response: Fe
          Df  Sum Sq Mean Sq F value    Pr(>F)
Site       3 134.222  44.741  89.883 1.679e-12 ***
Residuals 22  10.951   0.498
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

[8]Fun archaeological fact: pieces of broken glass are called “shards” while broken pottery pieces are “sherds.”

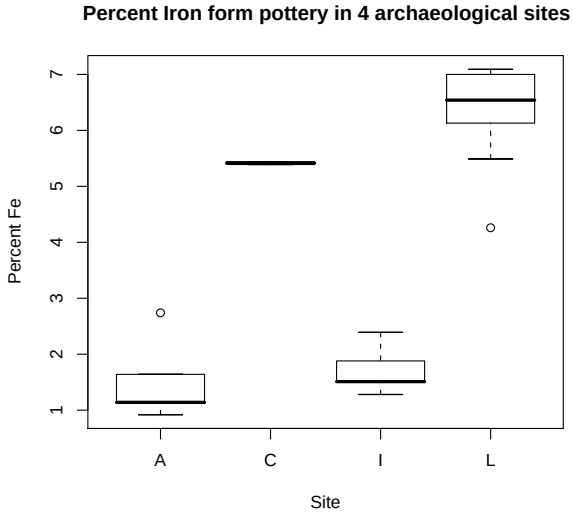


Figure 5.3: Date on percent iron from four archaeological sites in Roman-era Britain. Data from Tubbs et al 1980.

From these results, there appears to be ample reason to suspect the sites are different. The probability that the sites all have the same mean is less than 10^{-12}.

Note that the anova table was used here, rather than `summary()` command t-tests. The default coding for this model is `contr.treatment`, in which each site is compared to the first site (and the first is compared to zero). These are not particularly meaningful in this case (although see below for all pairwise comparisons). These anova results are interpretable, as is, only to say that at least one mean is different from the others.

Multiple comparisons

Often the question is not whether the simple null hypothesis (equality of all means) is true, but rather which are equal and which are different. Note that this is a bit tricky: one site might not be statistically different from second site, and the second site is not statistically different from the third, but the first *is* statistically different from the third.

A real concern when performing 6 tests (or in general, $\binom{l}{2}$ tests, where l is the number of levels), the more tests performed, the greater the chance that we get a type I error (rejecting the null when it is true). To get around this problem Tukey's "honest significant difference" (HSD) can be used. Rather than comparing each mean to a normal model, the difference between means is compared to the distribution of the maximum difference for observations from the same normal model. This this way, the p value output is, in fact, a proper p value.

To use Tukey's HSD, use the `TukeyHSD()` command. Unfortunately, it (at present) only takes `aov()` objects as an argument. For our purposes, `aov()` is just like `lm()`.

```
> met.aov<-aov(Fe~Site,data=met)
```

```
> TukeyHSD(met.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Fe ~ Site, data = met)

$Site
          diff        lwr       upr     p adj
C-A  3.9030000  2.2638764  5.542124 0.0000068
I-A  0.2000000 -1.0390609  1.439061 0.9692779
L-A  4.8601429  3.8394609  5.880825 0.0000000
I-C -3.7030000 -5.3421236 -2.063876 0.0000146
L-C  0.9571429 -0.5238182  2.438104 0.3023764
L-I  4.6601429  3.6394609  5.680825 0.0000000
```

From these data, it is perhaps apparent that site A is not significantly different (at the 0.05 level) from site I, while I is significantly different from site C, and C is not significantly different from site L.

5.8.3 Spices by country

The "spiciness" of different foods, as anyone who eats at places other than fast-food[9] can attest, varies greatly. Some people have suggested that spiciness may have an antimicrobial property. To examine this hypothesis, researchers gathered data on traditional foods from different parts of the world.[10] This anova will determine if the number of spices per recipe varies significantly by continent.

First, read in the data (there is a lot more in the table than just spice and continent). A box plot can be used to visually examine the relationship (see figure 5.4).

```
> spi<-read.table(
     file="http://students.washington.edu/nesse/qerm514/data/spices.txt",
     header=T)
> boxplot(spi$Meat~spi$cont)
```

An anova table can be computed to determine if there is significant variation in the spiciness.

```
> spi.lm<-lm(Meat~cont,data=spi)
> anova(spi.lm)
Analysis of Variance Table

Response: Meat
          Df Sum Sq Mean Sq F value Pr(>F)
cont       5 19.992   3.998  1.5606 0.1994
Residuals 32 81.987   2.562
```

[9] Does salt count as a spice?

[10] Published in Sherman, PW and Hash, GA. 2001. Why vegetable recipes are not very spicy. *Evolution and Human Behavior* 22: 147-163.

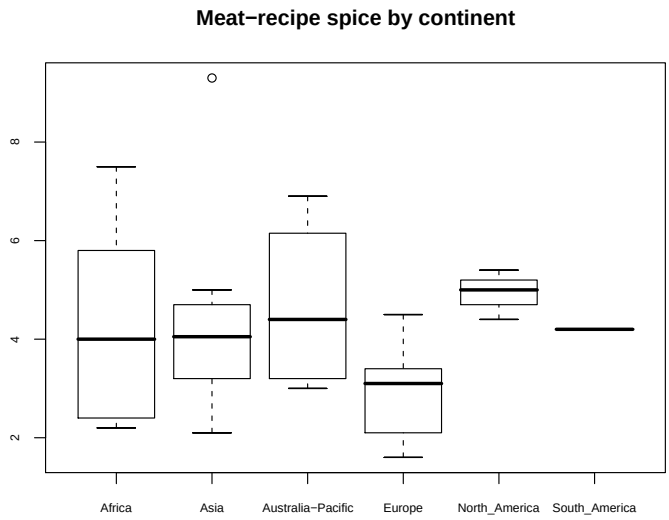


Figure 5.4: Average spiciness in meat-based recipes by continent. Note that there is only one observation in South America.

The model here seems insignificant, meaning continent is probably not a useful predictor of spiciness. Suppose the question were different, however. Suppose the researcher were based in North America, and wanted to know if the rest of the world was significantly different from North America. To test the differences between the rest of the world and North America, the treatment coding can be used. Ordinarily, the first level of the treatment coding is the base, however it can be changed as an argument of `contr.treatment()`

```
> spi$cont<-C(cont, contr = contr.treatment(n = 6, base = 5))
> spi.lm<-lm(Meat~cont,data=spi)
> summary(spi.lm)

Call:
lm(formula = Meat ~ cont, data = spi)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17000 -1.02214  0.01071  0.66042  5.03000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9333     0.9241   5.338 7.42e-06 ***
cont1        -0.6167     1.1318  -0.545   0.5896
```

```
cont2        -0.6633     1.0537  -0.630   0.5335
cont3        -0.2583     1.2225  -0.211   0.8340
cont4        -1.9548     1.0184  -1.920   0.0639 .
cont6        -0.7333     1.8483  -0.397   0.6942
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.601 on 32 degrees of freedom
Multiple R-Squared: 0.196,      Adjusted R-squared: 0.07042
F-statistic: 1.561 on 5 and 32 DF,  p-value: 0.1994
```

Unfortunately, in switching contrasts, R lost the names of the levels. Here the intercept represents the North American mean spiciness. The only one which comes close to being significantly different, however, is 4, which is Europe. Caution should be used here, however, since we're doing a bunch of tests, there is often one (even given random data) which comes up as significant or almost significant by chance alone.

Thus the conclusion here is probably that the spiciness does not vary significantly between continents.

Lecture 6

Two-way anova

6.1 Main ideas

- Multiple factor models
- μ notation for anova models
- Interactions
- Interaction plots

6.2 `R` functions

6.3 Multiple categorical predictors

The situation dealt with in lecture 5 was a single categorical predictor and continuous response. Of course, it is often the case that multiple predictors, both continuous and categorical, might be useful in a model. This lecture discusses models arising from multiple categorical predictors, while the next lecture on Ancova deals with mixed categorical and continuous predictors.

6.3.1 Models

Full model

The term **Full model** *in many text books is taken to be the same as* **saturated model**, *that is a model with the same number of terms as points being fit. This is* **not** *the definition here. For these notes, a full model is a model with all interaction terms, regardless of the number of points being fit. For more on saturated models, see section 13.3.2.*

The introduction of several categorical predictors gives several options for modeling. One straight-forward way of modeling a response is to fit a mean to each unique combination of predictors. For instance, if the first of two predictors had two levels and the second of the predictors had

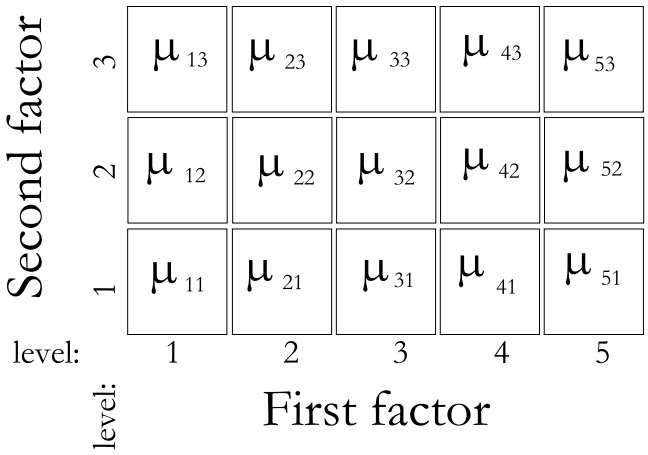


Figure 6.1: A full two-way anova, with 5 levels for the first factor and three for the second. Each unique combination of levels is fit as a (potentially) different mean. Since there are 15 such unique combinations for this model, 15 means must be fit.

three levels, this approach would fit six means. This model fits every unique combination of levels as independent.

The model equation, compare to equation 5.1, is fairly simple. Here i indexes the observation in group j and k (that is, the first group has level j and the second level k).

$$Y_{ijk} = \mu_{jk} + \epsilon_{ijk} \tag{6.1}$$

Graphically, for two levels, think of every possible combination of levels as a grid, with levels of the first factor on the horizontal and levels of the second factor on the vertical. This approach fits a mean to each cell. See figure 6.3.1.

Coding this model into a model matrix is discussed below.

Additive model

There are several possible issues which could arise using the full model with interactions. The number of fit variables grows with as the product of the levels, so three categorical predictors each with five levels would yield 125 fit parameters (plus a variance). As such, the data requirements to fit such a model would grow very quickly. It may also be reasonable to suspect that the different predictors do not have an interaction with each other.

A simpler model could be used; assuming there are no interactions between the factors, simply fit an additive term for each effect. Each level of each factor is associated with a fit parameter, bringing the total to $a + b$. A given response can be broken down into the sum of (i) a grand mean, (ii) the effect of being at level j of the first factor, and (iii) the effect of being at level k of the second

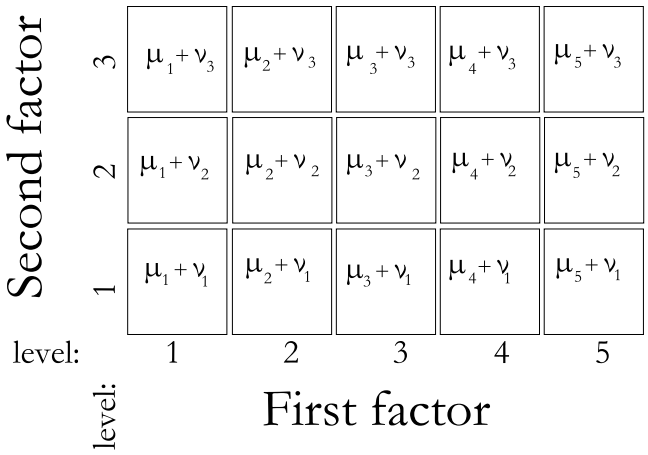


Figure 6.2: The additive model with two factors. The mean response in each cell is given by the sum of the effect of the first factor level and the sum of the second factor level.

factor. Each of those is a fit parameters, however unlike the full model, it does not depend on the other levels. Thus the effect of being at level j for the first factor is the same value no matter the level of the second factor.

For the two factor additive model, the equation looks a bit more complex than equation 6.1. In fact, however, it contains fewer parameters (since there are fewer means to estimate). In equation 6.1 there are ab means μ_{jk} to fit, one for each unique combination of j and k. In equation 6.2, by contrast, there are only a values for μ (one for each level of the first factor) and b for ν (one for each level of the second factor), bringing the total to $a + b$.

$$Y_{ijk} = \mu_j + \nu_k + \epsilon_{ijk} \tag{6.2}$$

As with the full model, a two-factor model can be thought of as a grid. Instead of fitting to a mean to each cell, however, values are fit to each level of each factor, rather than to each cell. The value of the response can be thought of as the sum of the effect of the first factor level and the second factor level. See figure 6.3.1

6.3.2 Three-plus way interactions

When three or more categorical predictors are used in a model, the full model has more than just all pairwise interactions. Three or more way interactions are also possible. This allows for models which include some interactions, but not all interactions. For instance, the model might include terms which have all pairwise interactions between the factors, but not the three way interactions. (Although it is possible to include models with interaction terms for which the main predictors are

not present, this is generally avoided, since the interpretation and inference is difficult and these models are considered unrealistic.)

An alternative way of writing equation 6.1 to look more in the form of equation 6.3.1 is to include terms like $(\mu\nu)_{jk}$, which are used to denote interactions between the first (μ) factor and the second (ν) factor. This can be a bit confusing, perhaps, as there are fit parameters $(\mu\nu)_{jk}$ for only some of the level combinations jk. Never-the-less, it is a means of describing a model with interactions explicitly shown.

$$y_{ijk} = \mu_j + \nu_k + (\mu\nu)_{jk} + \epsilon_{ijk} \tag{6.3}$$

The notation of multiplying the means to show an interaction is, it turns out, more than convention. When coding interactions, they are coded exactly as the (pointwise)[1] products of columns coding levels of the interacting factors. See below. Note that in equation 6.3, the $(\mu\nu)_{jk}$ term represents $ab - a - b$ fit parameters (including the mean).

It is possible to extend this model notation, to include three or more factors. Associate μ, ν, and η as the effect associated with each of three factors. The possible interaction terms, therefore, are $\mu\nu$, $\mu\eta$, $\nu\eta$, and $\mu\nu\eta$. This last one, the three way interaction term, is new. Any combination of these can be included, usually subject to the rule that the inclusion of higher order terms means lower order terms should also be included (thus if the three-way interaction term is included, all should be included under this rule).

6.3.3 Coding

A little bit of attention should be paid to how models with interactions are coded. The notation in 6.3 is perhaps most instructive. The model matrix, as always, starts with a column of 1s, to represent the mean. The two factors are coded using any of the coding options designated (the default for unordered factors is `contr.treatment`, see section sec:unorderedfactors). Thus if the first and second factors have a and b levels respectively, these additive terms require $a-1$ and $b-1$ columns to include them in the model. These columns of the model matrix correspond to dummy variables x_1 through x_a and w_1 though w_b.

Interactions are coded as products of the columns of the coding. An additional $(a-1)(b-1)$ columns are needed to code for a two-way interaction. Written as an equation, this becomes

$$y_{ijk} = \beta_0 + \sum_{t=1}^{a} \beta_t x_{jt} + \sum_{t=1}^{b} \beta_{a+t} w_{kt} \sum_{t=1}^{a} \sum_{r=1}^{b} \beta_{b(t-1)+t} x_{jt} w_{kr} + \epsilon_{ijk}. \tag{6.4}$$

Admittedly, equation 6.4 is not the prettiest equation around. Considering a simpler example, however, make make it easier to follow. Suppose two factors μ and ν are used as predictors with y as a response. If μ has two levels, and ν has three, then the additive terms will take 1 and 2 parameters respectively. The full model, with the interaction terms, is shown below.

$$y_{ijk} = \beta_0 + \underbrace{\beta_1 x_1}_{\text{First factor}} + \underbrace{\beta_2 w_1 + \beta_3 w_2}_{\text{Second factor}} + \underbrace{\beta_4 x_1 w_1 + \beta_5 x 1_w 2}_{\text{Interaction terms}} + \epsilon_{ijk} \tag{6.5}$$

[1] Pointwise product of vectors is to multiply each component by the component in the corresponding position. In more formal settings, this sort of matrix product is called the "Hadamard product." Often people beginning in matrix algebra naively multiply matrices using the Hadamard product, only to be corrected to ordinary matrix multiplication.

6.4 Inference

Although the coding is more complex, two way anova is still a linear model, set up in the usual way: $y = X\beta + \epsilon$. Estimates of β are still obtained using $\hat{\beta} = (X'X)^{-1}X'y$, and those parameters can still be assessed using the t test. Due to the coding complexities, however, little attention is paid here to the t test values. Instead, emphasis is put on the F test.

The reduction the sum of squared due to the addition of another factor can, as with the other linear models discussed so far, be assessed using an F test. In addition to the effect of each factor, the reduction in sum of squares due to the inclusion of interactions can also be assessed (if there are enough data to fit the full model).

6.4.1 F tests and t tests

As with all of multiple regression, we still have F tests and t tests. As with one way anova, t tests depend on coding, while F tests do not. Recall that for multiple regressions, the order of terms mattered for partial F tests. In the special case of balanced anovas (the same number of responses in each combination of factor levels), however, order no longer matters.[2]

6.4.2 Interaction F-test

The number of fit parameters for the full model is ab, for a levels of the first factor and b levels for the second. The additive model has $a + b$ parameters. Since these are nested models (all additive terms are in the full model) it is possible to test the significance of the interaction terms using an F test. Let n be the number of total observations.

$$\frac{(\mathrm{RSS}_{\mathrm{Add}} - \mathrm{RSS}_{\mathrm{Full}})/((a-1)(b-1))}{\mathrm{RSS}_{\mathrm{Full}}/(n-ab)} \sim F_{(a-1)(b-1),n-ab} \tag{6.6}$$

Here $\mathrm{RSS}_{\mathrm{Add}}$ is the additive model residual sum of squares, and $\mathrm{RSS}_{\mathrm{Full}}$ is the full model residual sum of squares.

6.5 Interaction plots

Although the F test can be an indicator of whether the interactions are significant, it is difficult perhaps to get an idea of what significant interactions actually mean. Developing a plot which shows interactions may not only shed light on whether to include interactions, but may also give some indication of what interactions mean.

Pairwise interactions can be plotted by using (cleverly titled) "interaction plots." The means response in each cell (see figure 6.3.1) is plotted against one of the predictors. The responses associated with the other predictor are connected by straight lines. See figure 6.5

If there are no interactions, that is under the additive model, the effect of being at a given level is the same, no matter the level of the other factor(s). Thus, the interaction plot should, under the additive model, produce parallel lines. The distance between the lines at the first factor level should be the same as the distance at the second level, etc. If the interaction plot lines cross or are distinctly not parallel, it is reasonable to suspect interactions.

[2] I didn't immediately come up with a proof of this, so it can be done as an extra credit assignment.

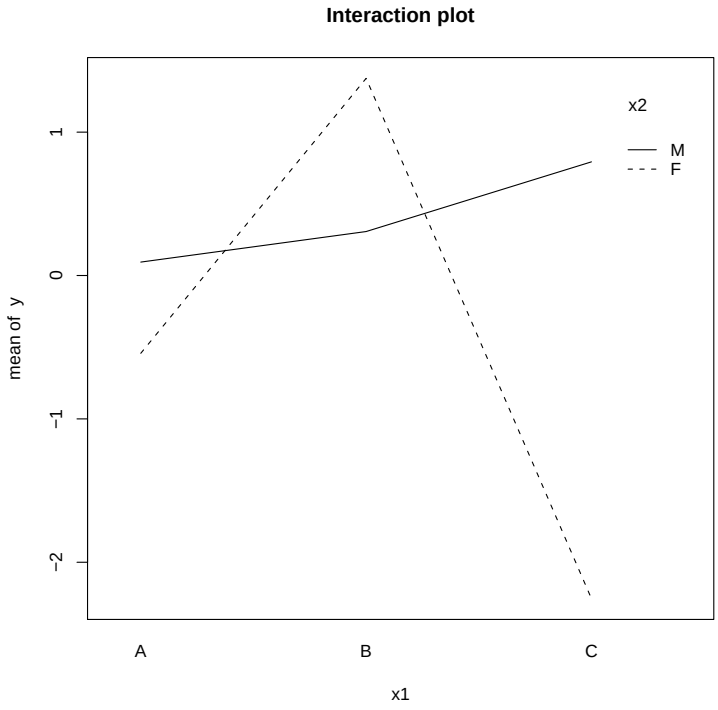


Figure 6.3: An example of an interaction plot (meaningless data), where two factors, the first with three levels and the second with two. The response is plotted on the vertical axis, against the first factor on the horizontal. The responses corresponding to the different levels of the second factor are connected. Since the interactions do not seem to be parallel here, it is likely that the interactions are present.

If the lines do cross, the level of one factor has an effect on the level of another factor. For instance (see section 6.6.1), consider a drug trial with sex (male or female) and treatment (drug A or drug B) as the predictors. Suppose the response, time to heal, is continuous and an anova model is used. If the interaction plot is crossed (see figure 6.6.1), so that the response for men on drug A is higher than women, and the response for women is higher for drug B than men. Thus the response of the drug is somewhat muddled–the response of the drug depends on sex of the patient.

When evaluating the drug, what experimenters would like to hear may be something simple like "drug A improves healing time by about t days." Absent an interaction, that sort of interpretation is possible. However if there is an interaction present, this sort of simple statement is not possible. Instead, the best that the data can indicate is "drug A changes healing time by t_1 for men over drug B, and changes by t_2 for women." This may be sufficient, however if there are many significant interactions, these statements will be increasingly complex.

6.6 Examples

Example	Section	Purpose
Drug interactions	6.6.1	Two-way anova with interactions
PCBs in Steller sea lions	6.6.2	Two way anova as an approach to spatial analysis.
Buttermilk	6.6.3	A two-way balenced anova.

6.6.1 Drug interactions

Suppose a drug company is conducting a trial of a new drug, called drug B. It is going to be compared to the standard treatment, drug A. The response being measured is the recovery time from acute statisticsitis (a dread disease). It is being tested in both men and women.

The responses for the thirty patients in the experiment are shown in the table below.

Drug	Sex	Recovery time
A	M	8.682842
A	F	2.927385
A	M	8.693901
A	F	2.388467
A	M	9.639283
A	F	3.317379
A	M	6.931770
A	F	3.959931
A	M	6.157346
A	F	4.526164
A	M	7.295531
A	F	2.384378
A	M	8.309264
A	F	3.563025
A	M	6.731654
B	F	8.054840
B	M	4.109288
B	F	8.691786
B	M	8.451489
B	F	7.484822
B	M	2.402233
B	F	6.518849
B	M	4.001061
B	F	7.901676
B	M	4.197783
B	F	9.782038
B	M	2.214755
B	F	9.227358
B	M	3.591897
B	F	5.462376

First, read in the data.

```
> drg <- read.table('http://students.washington.edu/nesse/qerm514/data/druganova.txt', header=T)
```

The full model can be specified in several ways in R. The expressions

```
> drg.sat.lm<-lm(res ~ sex + drug + sex:drug, data = drg)
```

or

```
> drg.sat.lm<-lm(res ~ sex * drug, data = drg)
```

both fit the same model. Note that the colon is used to denote an interaction, while the asterisk is used to denote the interaction of the factors, and all lower level interactions (and the factors themselves).

Now generate an anova table for the data.

```
> anova(drg.sat.lm)
Analysis of Variance Table

Response: res
          Df  Sum Sq Mean Sq F value    Pr(>F)
sex        1   0.908   0.908  0.4413    0.5123
drug       1   1.609   1.609  0.7819    0.3847
sex:drug   1 127.422 127.422 61.9242 2.406e-08 ***
Residuals 26  53.500   2.058
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Note here that the interaction is highly significant, while the factors themselves are not. If, instead of the full model, this investigation had started with the additive model, it might have seemed initially reasonable to reject both sex and drug as useful predictors[3]. The full model, however, shows highly significant interactions, and the general rule is if there are significant interactions, the factors themselves (and lower order interactions) should be kept.

To see what is going on here, examine the interactions plot.

```
> interaction.plot(drg$sex, drg$drug, response = drg$res, main ="Interaction plot")
```

This outputs figure 6.6.1. Since the plot lines are not parallel, it is reasonable to suspect interactions. In fact, the response for men and women are nearly opposite. How well the new drug works best depends on the sex of the patient.

The responses shown in 6.6.1 are also good indicators of why the additive model is non-significant. Because of the near-opposite reactions of men and women to the drugs, either of the drugs can not be said to have a single additive effect. Thus the fit additive effect of each drug is very near zero.

6.6.2 PCBs in Steller sea lions

Steller sea lions (*Eumetopias jubatus*) are a large marine mammal which live in the North Pacific. They are managed as two distinct population segments[4]: The western stock, spanning the bulk of Alaska into the Sea of Okhotsk. and the Kuril Islands, sometimes straying as far south as Japan, and the eastern stock, comprising most of the US and Canadian Pacific coasts, and a portion of Alaska. The western stock is currently classified as endangered under the Endangered Species Act.

The cause of the decline of the western stock is not well established, although there is a substantial amount of work being done to determine a cause. One suggested problem is the contamination of polychlorinated biphenyls (PCBs) and other organochlorides. To look at possible patterns of contamination, data were collected from blubber samples from pups at six locations across the range.[5] Data were collected from both males and females.

[3]Maybe rock and roll would help.

[4]Populations which have so little interaction they can be managed separately for legal purposes such as the endangered species act.

[5]These are not actual data, but were generated to reflect the pattern reported in Myers et al 2008.

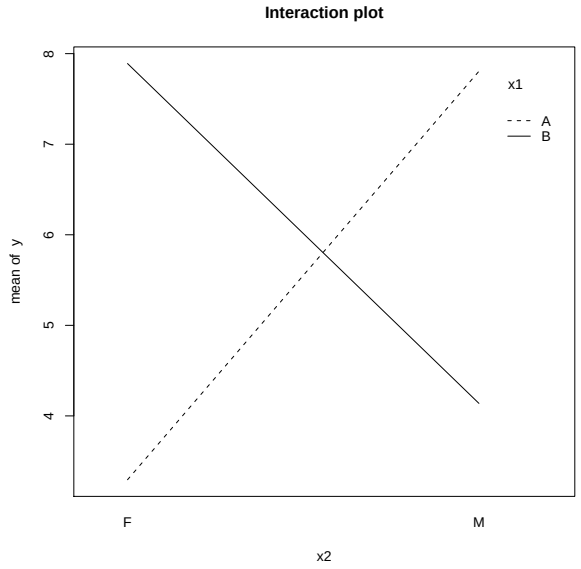


Figure 6.4: An interaction plot from the example in section 6.6.1. The crossing lines give an indication that there is an interaction. Note that the higher response for men is using the opposite drug as it is for women.

PCB (ng/g wet weight)	Sex	Location	PCB (ng/g wet weight)	Sex	Location
1.9747603	M	amak	1.8514759	M	marmot
2.3046883	M	amak	1.0082548	M	marmot
1.8506436	M	amak	2.7451632	M	marmot
2.1271506	M	amak	1.3363535	M	marmot
2.2513681	M	amak	1.4685980	M	marmot
6.3069299	F	amak	1.7650534	M	marmot
6.3365125	F	amak	0.7905333	M	marmot
4.3622643	F	amak	0.4580265	M	marmot
7.1926219	F	amak	1.2193069	M	marmot
4.6556592	F	amak	3.8454851	F	marmot
1.7577994	M	pinnacle	1.5152488	F	marmot
2.8752499	M	pinnacle	0.4677070	F	marmot
1.5281192	M	pinnacle	4.3012527	F	marmot
1.6349244	M	pinnacle	0.2718075	F	marmot
2.7095608	M	pinnacle	0.1999261	F	marmot
1.6169130	M	pinnacle	3.3727760	F	marmot
2.8457950	F	pinnacle	3.1795057	F	marmot
3.3852219	F	pinnacle	0.7882853	F	marmot
4.9387985	F	pinnacle	2.1294914	M	sugarloaf
1.3941679	M	atkins	0.6086950	M	sugarloaf
1.4932371	M	atkins	3.3245745	M	sugarloaf
0.9737255	M	atkins	2.3244367	M	sugarloaf
1.2114022	M	atkins	3.0269404	M	sugarloaf
0.9730735	F	atkins	0.9478122	M	sugarloaf
1.7449031	F	atkins	1.9052177	M	sugarloaf
1.6114389	F	atkins	1.9249998	M	sugarloaf
1.1519923	F	atkins	0.5235522	M	sugarloaf
0.8237885	F	atkins	0.4096965	M	sugarloaf
1.8689524	M	chirikof	2.5179964	M	sugarloaf
1.6794136	M	chirikof	1.5729853	F	sugarloaf
0.7945591	M	chirikof	0.8310353	F	sugarloaf
1.0285355	M	chirikof	0.9451559	F	sugarloaf
2.1312664	M	chirikof	1.2475075	F	sugarloaf
5.4391665	F	chirikof	1.0813220	F	sugarloaf
5.5678537	F	chirikof	2.0835720	F	sugarloaf
4.1549385	F	chirikof	0.9306772	F	sugarloaf
6.1351753	F	chirikof	2.0416800	F	sugarloaf
3.1011772	F	chirikof	2.0346890	F	sugarloaf

It is perhaps reasonable to determine if the PCB load carried by these pups varies by location and sex. To do this, a two-way anova model is employed. Since the question of interest is geographical, these geographical predictors have a natural order: longitudinal. First we'll map[6] the locations.[7] See figure 6.5.

```
> ss<-read.table(
      file="http://students.washington.edu/nesse/qerm514/data/pcbSSL.txt",
      header=T)
>
> # Load mapping libraries
> library(maps)
> library(mapdata)
```

[6]Fun fact about R: it can be used to draw and manipulate maps. Many GIS functions are available in the sp package.

[7]The location data is not in the dataset, but easily obtainable.

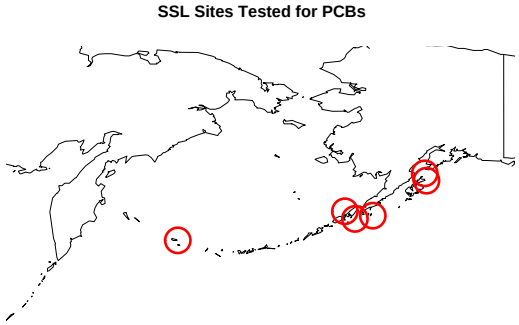


Figure 6.5: Steller sea lion rookeries where pups were tested for PCB levels. Note that the sites are roughly on an east-west line. Sites, going from west to east, are Chirikof, Amak, Pinnicle, Atkins, Sugarloaf, and Marmot.

```
>
> #Draw map
> map("world2Hires",xlim=c(150,220),ylim=c(30,70))
> points(c(196.80,208.20,200.70,198.20,207.96,173.45),
         c(55.46,58.23,55.06,54.77,58.89,52.83),col="red",
         cex=4,lwd=3)
> title(main="SSL Sites Tested for PCBs")
```

This suggests entering the sites as ordered factors–going from west to east: Chirikof, Amak, Pinnacle, Atkins, Sugarloaf, and Marmot. To do this, the `ordered()` command is used.

```
> ss$location<-ordered(ss$location,
    levels=c("chirikof","amak","pinnacle",
             "atkins","sugarloaf","marmot"))
```

On the question of interactions, an interaction plot would be useful.
6.6.

```
> interaction.plot(ss$sex,ss$location,response=ss$pcb,
      main="Steller sea lion interaction plot",
      xlab="Sex",
      ylab="PCB concentration (mean)")
```

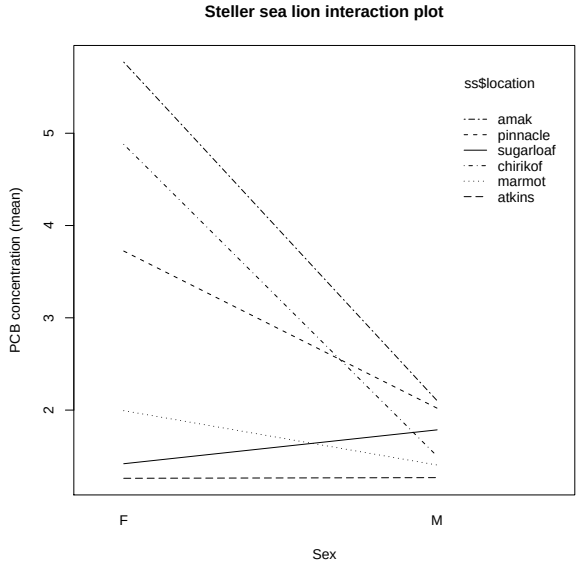


Figure 6.6: A plot to look at the interaction between sex and location for Steller sea lion contamination.

It does appear that there are some plausible arguments for interaction. Ideally, the model choice is driven not by plots like this, however, but by the science–do we have reason to look for an interaction, or expect an interaction will occur?

The model with interactions is fit below.

```
> ss.lm<-lm(pcb~location*sex,data=ss)
> anova(ss.lm)
Analysis of Variance Table

Response: pcb
             Df Sum Sq Mean Sq F value    Pr(>F)
location      5 60.376  12.075  13.599 4.844e-09 ***
sex           1 25.553  25.553  28.778 1.195e-06 ***
location:sex  5 44.675   8.935  10.063 3.872e-07 ***
Residuals    64 56.828   0.888
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The interaction term is indeed highly significant, indicating differences between males and females at different locations. The predictors themselves are also highly significant. Now consider the t tests (recall the default coding for ordered factors is `contr.poly`, indicating if there is a linear, quadratic, or other polynomial trend in the data.

```
> summary(ss.lm)
```

```
Call:
lm(formula = pcb ~ location * sex, data = ss)

Residuals:
     Min       1Q   Median       3Q      Max
-1.79363 -0.48074 -0.01565  0.56106  2.30770

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.17451    0.16882  18.804  < 2e-16 ***
location.L      -3.57958    0.37542  -9.535 6.66e-14 ***
location.Q       0.78987    0.41917   1.884  0.06406 .
location.C       1.92919    0.39454   4.890 7.12e-06 ***
location^4      -0.89327    0.40781  -2.190  0.03214 *
location^5      -0.36210    0.46521  -0.778  0.43923
sexM            -1.49429    0.23158  -6.453 1.68e-08 ***
location.L:sexM  3.31913    0.52735   6.294 3.16e-08 ***
location.Q:sexM -1.06431    0.57573  -1.849  0.06913 .
location.C:sexM -1.57575    0.54519  -2.890  0.00525 **
location^4:sexM  0.48128    0.55862   0.862  0.39215
location^5:sexM -0.01832    0.62447  -0.029  0.97669
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9423 on 64 degrees of freedom
Multiple R-Squared: 0.6968,     Adjusted R-squared: 0.6447
F-statistic: 13.37 on 11 and 64 DF,  p-value: 8.898e-13
```

Thus, reading the `location.L` line, it is evident there is a strong trend in the sites, with the heaviest concentration in the west (as evidenced by the negative slope). While admittedly this is a quick-and-dirty approach to spatial analysis (close points, we might guess are correlated, just as in the Leaning Tower of Pisa example, section 4.7.3, it is indeed illustrative of a real effect.

6.6.3 Butterfat in milk

These data come from the percent butterfat in five species of cows. Each cow is aged either 2 years or mature, and the breed is listed for each. This allows a two-way anova, with the breed as one predictor and age as another predictor. The response here is the percentage butterfat in the milk.

First, read in the data.

```
> bu<-read.table(
    file="http://students.washington.edu/nesse/qerm514/data/butter.txt",
    header=T)
```

Before jumping right to a model, it may be helpful to look at the data, first as boxplots, then using an interaction plot. The results are shown in figure 6.7

```
> boxplot(bu$Butterfat~bu$Age,main="Age")
> boxplot(bu$Butterfat~bu$Breed,main="Breed")
> interaction.plot(bu$Breed,bu$Age,response=bu$Butterfat)
```

The boxplots do appear to show some variation by predictor. The interaction plot, however, strongly indicates no interaction between the predictors. Thus the additive model is the correct choice here.

```
> bu.lm<-lm(Butterfat ~ Breed + Age,data=bu)
> summary(bu.lm)

Call:
lm(formula = Butterfat ~ Breed + Age, data = bu)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0202 -0.2373 -0.0640  0.2617  1.2098

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.00770    0.10135  39.541  < 2e-16 ***
BreedCanadian         0.37850    0.13085   2.893  0.00475 **
BreedGuernsey         0.89000    0.13085   6.802 9.48e-10 ***
BreedHolstein-Fresian -0.39050    0.13085  -2.984  0.00362 **
BreedJersey           1.23250    0.13085   9.419 3.16e-15 ***
AgeMature             0.10460    0.08276   1.264  0.20937
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4138 on 94 degrees of freedom
Multiple R-Squared: 0.6825,     Adjusted R-squared: 0.6656
F-statistic: 40.41 on 5 and 94 DF,  p-value: < 2.2e-16

> anova(bu.lm)
Analysis of Variance Table

Response: Butterfat
          Df Sum Sq Mean Sq F value Pr(>F)
Breed      4 34.321   8.580 50.1150 <2e-16 ***
Age        1  0.274   0.274  1.5976 0.2094
Residuals 94 16.094   0.171
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Here, each of the breeds appears to be significant. Since this is a balenced design, the p value of the F test is order independent.

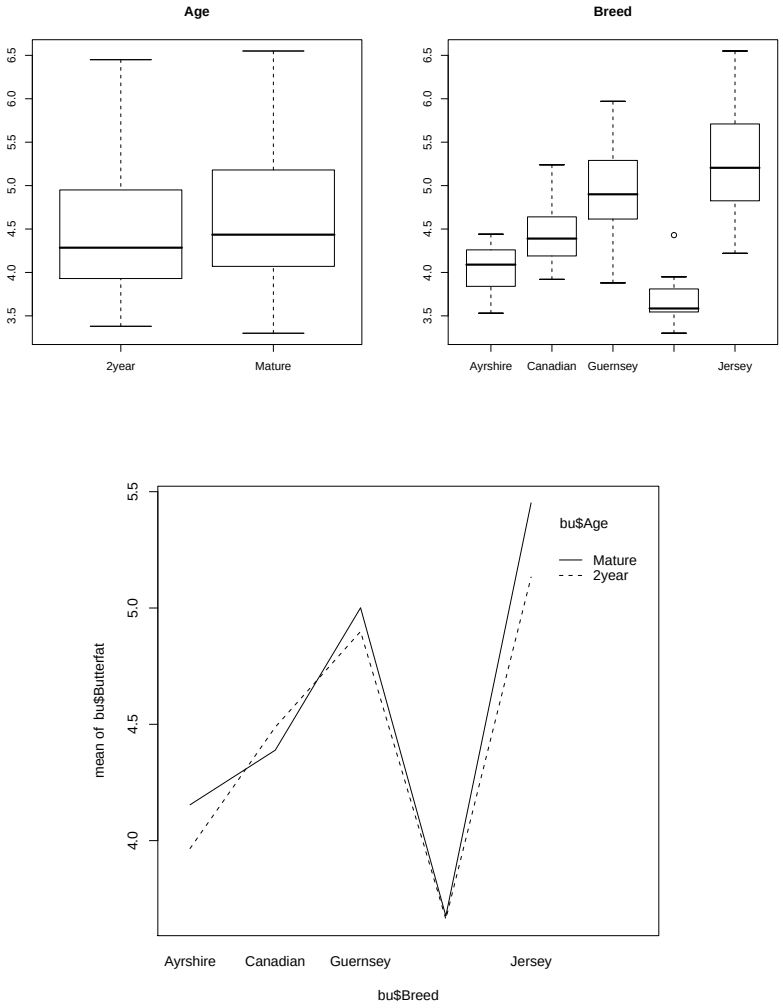


Figure 6.7: Boxplots and interaction plot for the butterfat data.

Lecture 7

Ancova

7.1 Main ideas

- Mixed categorical and continuous predictors
  - Interpretation of models
  - Plotting

7.2 Mixed predictors

The course started with continuous predictors and continuous response. Next categorical predictors were considered, again with a continuous response. The natural progression is to consider models with both continuous and categorical predictors in the same model. There are no new inferential techniques–the F and t tests are still where it is at–but the interpretation is a bit new and some multi-level plotting may be useful.

Mixed continuous and categorical predictors enter the design matrix as they do when doing ordinary regression. Categorical predictors are coded, while continuous predictors are single columns. The coding of categorical predictors, as it is for one-way anova, is important for the interpretation of results.

7.2.1 Interactions

When dealing with two-way anova, the possibility of interactions arose. In fact, interactions can be defined for continuous predictors as well, although it is less common to do so. The possibility arises, however, for interactions between categorical and continuous predictors. These are the product of the coded variable(s) for the categorical predictor and the continuous predictor.

For example, a categorical predictor with three levels (coded into the dummy variables x_1 and x_2) and a continuous variable coded with w would be written

$$y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 w + \beta_4 x_1 w + \beta_5 x_2 w + \epsilon_{ij} \tag{7.1}$$

The β_4 and β_5 terms enter the model coding for the interactions.

7.3 Interpretation

7.3.1 Simple example

The ancova matrix equation looks identical to all the others seen thus far, $Y = \beta X + \epsilon$. Written out in univariate form, however, more interpretation is possible. The simplest model, a single continuous predictor and a single categorical predictor with two levels, is shown in equation 7.2. The continuous predictor is coded with a w and the categorical predictor is coded with an x.

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \beta_3 x_i w_i + \epsilon_i \tag{7.2}$$

The default coding in `R` for unordered factors would represents $x = 0$ for the first level, and $x = 1$ for the second level. Thus the line

$$y_i = \beta_0 + \beta_1 w_i + \epsilon_i$$

is being fit to the observations in the first level, while

$$y_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) w_i + \epsilon_i$$

is fit to observations with the second level of the categorical predictor. This could be seen as fitting a different mean and intercept to each group; in a sense, a completely different line is fit to each observation[1]

The parameters under this coding (interpretation is always coding dependent) is β_0 is the intercept of the first level's line, while β_2 is the difference of the second level intercept and the first. Likewise, β_1 is the slope of the line going though the first level data, and β_3 is the difference between the second level slope and the first.

A model without the interaction term is fitting two lines with the same slope, but different intercepts (see figure 7.3.1 for example).

7.3.2 Interpretation using treatment coding

Recall that if a categorical predictor has k levels, it will have $k - 1$ coded dummy variables. In the treatment coding, as described in table 5.1, the first level is all zeros, the second level is all zeros except the first dummy variables, the third level is zeros except for the second dummy variable, and so on. This coding is the same for ordinary one-way anova and ancova.

Thus, if the model under consideration is the additive model, with a continuous predictor x and categorical dummy variables $w_1, \ldots, w_{k-1}$, the expression is as shown in equation 7.3.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_{i1} + \cdots + \beta_k w_{i(k-1)} + \epsilon_i \tag{7.3}$$

Under the treatment coding for w, each $w_{ij} = 1$ if and only if observation i is at level $j + 1$ of the categorical predictor (and is otherwise zero)[2]. Thus the intercept for the line fit to observations which are at level j is $\beta_0 + \beta_j$, unless $j = 1$, in which case, the intercept is simply β_0. The slope, under the additive model, is the same for all lines, and is β_1.

[1] In a sense because the model uses all the data to estimate the common variance, $\hat{\sigma}^2$.

[2] This is not a deep statement; it is an arbitrary choice of the treatment coding. See table 5.1.

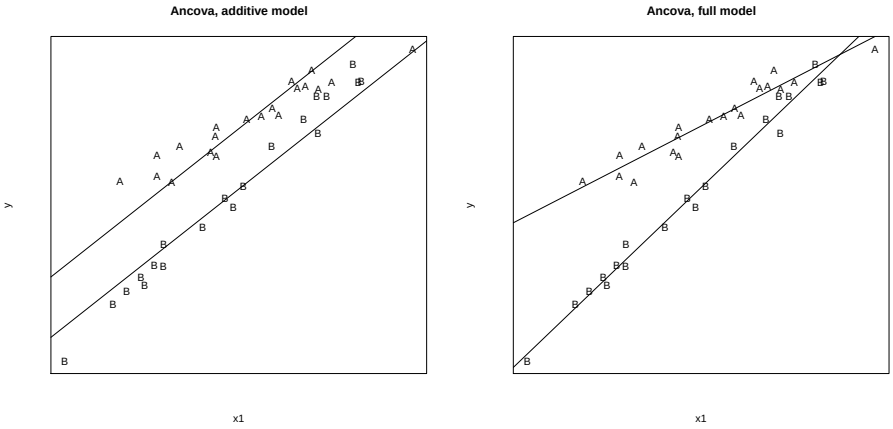


Figure 7.1: A continuous response with one continuous and one categorical predictor. The categorical predictor has two levels, `A` and `B`. The left diagram shows the additive model, $y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \epsilon_i$, where the data share the same slope, but have different intercepts. The right diagram shows the full model, in which data from different levels can have different slopes as well as intercepts.

Alternatively, if the model under consideration is the full model, and an interaction between the categorical and continuous variable is included, the model becomes equation 7.4.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_{i1} + \cdots + \beta_{k-1} w_{i(k-1)} + \beta_k w_{i1} x_i + \cdots + \beta_{2k-1} w_{i(k-1)} x_i + \epsilon_i \tag{7.4}$$

In this case, the intercept for an observation at level j of the categorical variable is as before, $\beta_0 + \beta_j$, unless $j = 1$. The slope, however, now also varies by factor level. For observations from the first level, the slope is β_1. However, for any other level j, the slope is $\beta_1 + \beta_{k+j-1}$.

7.4 Examples

7.4.1 Simple ancova

The data described below are a continuous response y with continuous predictor x_1 and categorical predictor x_2. The task is to fit these data using (i) additive effects only (no interaction), (ii) interactions only (no additive effect) and (iii) the full model with both additive effects and an interactions.

y	x1	x2	y	x1	x2
31.3348627	13.2352840	A	8.9016658	0.2209877	B
18.2370882	0.1014272	A	16.1464108	10.5872125	B
12.8339293	-5.7342572	A	6.8182943	-0.6297167	B
-3.1417008	-20.7883531	A	-17.8312727	-17.7780674	B
17.4008356	-1.2918407	A	15.2856504	5.4844520	B
18.3083353	-4.3177938	A	13.6167815	3.0262780	B
11.4659274	-10.0120084	A	18.4175860	9.4466437	B
20.4996640	-1.6091636	A	10.6130354	3.8535721	B
26.3937454	7.4013941	A	37.1863752	22.1139841	B
13.4741328	-8.1727823	A	-4.9477160	-9.0291949	B
26.1189077	4.8634912	A	25.1400266	-15.0223254	C
24.5608564	3.2204310	A	22.9708543	-10.8384638	C
16.6855968	-3.1286942	A	32.8890049	1.6786049	C
16.4898827	-1.7034685	A	39.0628689	21.2108532	C
21.6210290	4.6151864	A	27.0142555	-9.2901033	C
35.2888125	14.0865760	A	33.4992281	6.2084796	C
25.9691849	12.2663179	A	30.5493851	2.7524899	C
34.8259888	15.8461491	A	40.7634197	9.9399694	C
32.5326804	10.9369327	A	31.6572733	4.7377756	C
-3.7979710	-19.2567810	A	29.9071940	-1.4192277	C
12.1980436	7.3487777	B	20.0563385	-17.1242166	C
-5.0857574	-7.9685668	B	34.8585338	9.8220896	C
-2.0421084	-2.7860629	B	27.7766427	-3.1505617	C
8.9811369	-2.4076084	B	34.8255668	3.6537993	C
19.9277150	11.7931269	B	24.3387150	-4.0085861	C
-8.2179242	-9.5975952	B	25.1004869	-8.4328551	C
-11.4621802	-13.9569273	B	33.6037363	6.2697188	C
-0.4601844	-6.1447577	B	28.8362954	-2.9736510	C
1.1417446	-2.2003800	B	26.9931437	-12.7113590	C
-5.3552249	-8.7177939	B	19.0588505	-13.0741432	C

First, read in the data.

```
ancEG<-read.table(file='http://students.washington.edu/nesse/qerm514/data/ancovaEG1.txt',header=T)
```

Now to plot the data; plots shown in figure 7.4.1

```
## No interactions
plot(x1,y,pch=as.character(x2),main="Ancova, no interaction")
temp.coef<-lm(y~x1+x2)$coeff
abline(a=temp.coef[1],b=temp.coef[2])
abline(a=temp.coef[1]+temp.coef[3],b=temp.coef[2])
abline(a=temp.coef[1]+temp.coef[4],b=temp.coef[2])

## Only interactions, no additive term
plot(x1,y,pch=as.character(x2),main="Ancova, only interaction")
temp.coef<-lm(y~x1+x1:x2)$coeff
abline(a=temp.coef[1],b=temp.coef[2])
abline(a=temp.coef[1],b=temp.coef[2]+temp.coef[3])
abline(a=temp.coef[1],b=temp.coef[2]+temp.coef[4])

## All interactions
plot(x1,y,pch=as.character(x2),main="Ancova, with interaction")
temp.coef<-lm(y~x1*x2)$coeff
```

```
abline(a=temp.coef[1],b=temp.coef[2])
abline(a=temp.coef[1]+temp.coef[3],b=temp.coef[2]+temp.coef[5])
abline(a=temp.coef[1]+temp.coef[4],b=temp.coef[2]+temp.coef[6])
```

The three models being fit are below: equation 7.5 is the additive model (same slope, different intercepts), 7.6 is the interaction only (different slopes, same intercept), and 7.7 is the full model. Recall that since x_2 has three levels, it must be coded with two dummy variables, written here as w_1 and w_2.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 w_{i1} + \beta_3 w_{i2} + \epsilon_i \quad (7.5)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 w_{i1} x_{i1} + \beta_3 w_{i2} x_{i1} + \epsilon_i \quad (7.6)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 w_{i1} + \beta_3 w_{i2} + \beta_4 w_{i1} x_{i1} + \beta_5 w_{i2} x_{i1} + \epsilon_i \quad (7.7)$$

7.4.2 1984 Presidential Election

In the 1984 presidential election, Ronald Regan (Republican incumbent) and Walter Mondale (Democrat, Vice president under Carter) faced off. States often consistently vote for the same party in subsequent elections, so it is reasonable to suspect that the percent of votes in 1980 for the Democratic candidate would be a good predictor of the 1984 outcome. Some people have suggested, however, that regions shifted their support in geographically cohesive ways. This view is borne out to some degree in figure 7.3.

First read in and plot the data.

```
> elec <- read.table(
      file="http://students.washington.edu/nesse/qerm514/data/84election.txt",
      header=T)
>
> plot(elec$Dem1980,elec$Dem1984,pch=as.numeric(elec$state.region),
      main="Percent of state voting Mondale in 1984",
      xlab="Percent voting Carter in 1980",
      ylab="Percent voting Mondale in 1984")
> legend(20,45,legend=levels(elec$state.region),pch=1:4)
```

To assess whether regions behaved differently in 1984 versus 1980, an analysis of covariance model can be used. It is not clear if it is reasonable to expect interaction terms or not based on the model description. They are included here to examine any effect. The four regions are `North Central`, `Northeast`, `South` and `West` (the default categorization in `state.region` which is a dataset built into `R`).

The model is set up and general descriptions are output.

```
> elec.lm<-lm(Dem1984 ~ Dem1980 * state.region,data=elec)
> anova(elec.lm)
Analysis of Variance Table

Response: Dem1984
                    Df Sum Sq Mean Sq F value    Pr(>F)
Dem1980              1 822.29  822.29 94.3856 2.639e-12 ***
state.region         3 413.70  137.90 15.8287 4.937e-07 ***
```

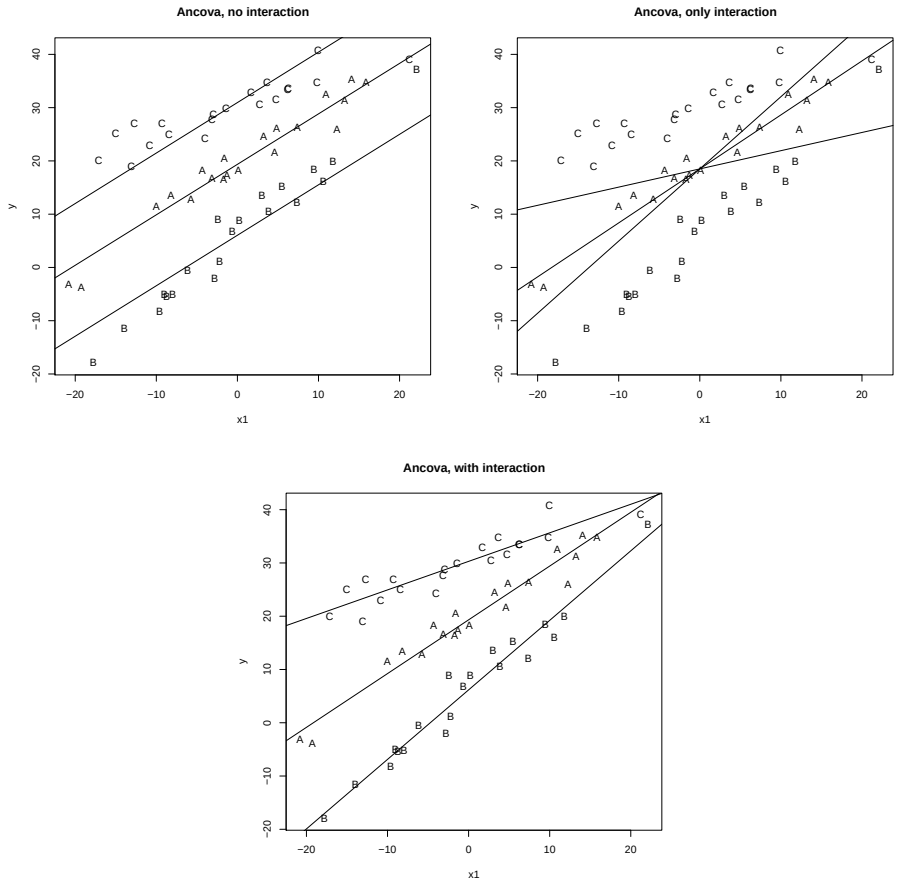


Figure 7.2: Three model for the same data. The first fits three lines with the same slope but different intercepts (no interactions, the categorical variable enters as an additive effect), the second shows a model with the same intercept but different slopes corresponding to interactions but no additive effect for the categorical predictor. The third is the full model, with both interactions and additive effects, fits a line with different slope and intercept to each level of the categorical variable.

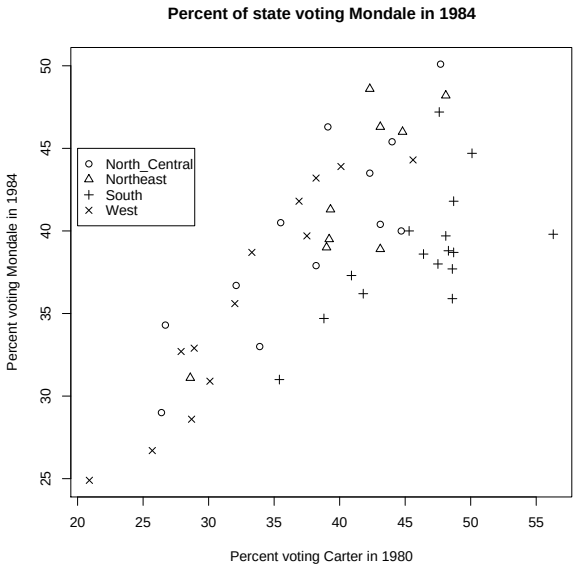


Figure 7.3: The percent of each state voting for Walter Mondale in 1984 plotted against the percentage voting for Carter in 1980. Different symbols are used to represent the region of the United States where each state is located.

```
Dem1980:state.region  3  56.66   18.89  2.1679    0.1060
Residuals            42 365.91    8.71
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

As expected, the vote in 1980 is a significant predictor of the vote in 1984. Moreover, looking at the anova table, there is a significant additive factor effect for the region, however the interaction term was not significant. An additive effect here seems to indicate regions changed how they voted in the election, above and beyond what would be expected in based only on the past election. However the lack of interaction significance does seem to indicate that the relation between the last election and the 1984 election itself did not change by region.

In more concrete terms, the expected percentage voting Democratic in the 1984 election was $\beta_0 + \beta_1 V_{1980} + \mu_{region}$. A linear term accounts for the effect of the 1980 percentage, V_{1980}, and a constant fit to each region. Geometrically, this is fitting a different line to each region, but with the same slope in all regions.

Looking now at the summary table.

```
> summary(elec.lm)

Call:
lm(formula = Dem1984 ~ Dem1980 * state.region, data = elec)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2611 -2.1408  0.1125  1.7044  7.8378

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    12.3090     4.8755   2.525   0.0154 *
Dem1980                         0.7260     0.1270   5.718 1.01e-06 ***
state.regionNortheast          -7.3383     9.2476  -0.794   0.4319
state.regionSouth               4.5403     8.6318   0.526   0.6017
state.regionWest               -7.8378     6.4579  -1.214   0.2317
Dem1980:state.regionNortheast   0.1833     0.2293   0.799   0.4286
Dem1980:state.regionSouth      -0.2531     0.1988  -1.273   0.2100
Dem1980:state.regionWest        0.2270     0.1795   1.265   0.2130
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.952 on 42 degrees of freedom
Multiple R-Squared: 0.7794,     Adjusted R-squared: 0.7426
F-statistic:  21.2 on 7 and 42 DF,  p-value: 6.764e-12
```

It may come as a surprise here that none of the coefficients for region are significant. Recall that the interpretation here depends on coding. Since the treatment coding was used, with `North Central` being the first (control) factor, (the default in `R`) these are meaningful as shown in table 7.1.

Name	Estimate	Meaning
(Intercept)	12.3090	The intercept of a line fit through the North Central data
Dem1980	0.7260	The slope of the line fit through the North Central data
state.regionNortheast	-7.3383	The difference between the intercept of the Northeast data and the North Central data.
state.regionSouth	4.5403	The difference between the intercept of the South data and the North Central data.
state.regionWest	-7.8378	The difference between the intercept of the West data and the North Central data.
Dem1980:state.regionNortheast	0.1833	The difference between the slope of the Northeast data and the North Central data.
Dem1980:state.regionSouth	-0.2531	The difference between the slope of the South data and the North Central data.
Dem1980:state.regionWest	0.2270	The difference between the slope of the West data and the North Central data.

Table 7.1: Interpretations of the summary table for the 1984 presidential election votes by state.

From these data it does appear that regions behaved differently in the 1984 election, above and beyond what might be expected from the 1980 election.

7.4.3 Tooth growth in Guinea Pigs

Nothing says fun like guinea pig teeth. In particular, what is the effect of Vitamin C on tooth growth in guinea pigs?[3] These data are from R and do not need to be imported.

len	supp	dose	len	supp	dose	len	supp	dose	len	supp	dose
4.2	VC	0.5	15.2	OJ	0.5	17.3	VC	1.0	25.2	OJ	1.0
11.5	VC	0.5	21.5	OJ	0.5	13.6	VC	1.0	25.8	OJ	1.0
7.3	VC	0.5	17.6	OJ	0.5	14.5	VC	1.0	21.2	OJ	1.0
5.8	VC	0.5	9.7	OJ	0.5	18.8	VC	1.0	14.5	OJ	1.0
6.4	VC	0.5	14.5	OJ	0.5	15.5	VC	1.0	27.3	OJ	1.0
10.0	VC	0.5	10.0	OJ	0.5	23.6	VC	2.0	25.5	OJ	2.0
11.2	VC	0.5	8.2	OJ	0.5	18.5	VC	2.0	26.4	OJ	2.0
11.2	VC	0.5	9.4	OJ	0.5	33.9	VC	2.0	22.4	OJ	2.0
5.2	VC	0.5	16.5	OJ	0.5	25.5	VC	2.0	24.5	OJ	2.0
7.0	VC	0.5	9.7	OJ	0.5	26.4	VC	2.0	24.8	OJ	2.0
16.5	VC	1.0	19.7	OJ	1.0	32.5	VC	2.0	30.9	OJ	2.0
16.5	VC	1.0	23.3	OJ	1.0	26.7	VC	2.0	26.4	OJ	2.0
15.2	VC	1.0	23.6	OJ	1.0	21.5	VC	2.0	27.3	OJ	2.0
17.3	VC	1.0	26.4	OJ	1.0	23.3	VC	2.0	29.4	OJ	2.0
22.5	VC	1.0	20.0	OJ	1.0	29.5	VC	2.0	23.0	OJ	2.0

The length of the odontoblast (a portion of tooth-growing tissue) is measured in guinea pigs as a function of two variables: dose of the Vitamin C (mg) and delivery method (plain (VC) or with orange juice (OJ)).

These data can be modeled in two ways: first as an anova model (dose as a continuous predictor, delivery as a categorical), and second as a two-way anova (both predictors as categorical variables).

First, the ancova model with all interactions.

```
> tg.ancova<-lm(len~supp*dose,data=ToothGrowth)
> anova(tg.ancova)
Analysis of Variance Table

Response: len
          Df  Sum Sq Mean Sq  F value    Pr(>F)
supp       1  205.35  205.35  12.3170 0.0008936 ***
dose       1 2224.30 2224.30 133.4151 < 2.2e-16 ***
supp:dose  1   88.92   88.92   5.3335 0.0246314 *
Residuals 56  933.63   16.67
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> summary(tg.ancova)

Call:
lm(formula = len ~ supp * dose, data = ToothGrowth)
```

[3] Guinea pigs, like humans, require vitamin C in their diet. Most other mammals (and reptiles and birds) can synthesize Vitamin C from glucose. Thus guinea pigs are often the animal model used for experiments in Vitamin C deficiency (scurvy).

```
Residuals:
     Min       1Q   Median       3Q      Max
-8.22643 -2.84625  0.05036  2.28929  7.93857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.550      1.581   7.304 1.09e-09 ***
suppVC        -8.255      2.236  -3.691 0.000507 ***
dose           7.811      1.195   6.534 2.03e-08 ***
suppVC:dose    3.904      1.691   2.309 0.024631 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.083 on 56 degrees of freedom
Multiple R-Squared: 0.7296,     Adjusted R-squared: 0.7151
F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```

These results show the slope and intercepts of each group vary significantly between the two delivery methods. Turn now to the two-way anova model, where dose enters as an ordered categorical variable.

```
> tg.anova<-lm(len~supp*ordered(dose,levels=c(0.5,1.0,2.0)),data=ToothGrowth)
> anova(tg.anova)
Analysis of Variance Table
Response: len
                                          Df  Sum Sq Mean Sq F value    Pr(>F)
supp                                       1  205.35  205.35  15.572 0.0002312 ***
ordered(dose, levels = c(0.5, 1, 2))       2 2426.43 1213.22  92.000 < 2.2e-16 ***
supp:ordered(dose, levels = c(0.5, 1, 2))  2  108.32   54.16   4.107 0.0218603 *
Residuals                                 54  712.11   13.19
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
>
> summary(tg.anova)

Call:
lm(formula = len ~ supp * ordered(dose, levels = c(0.5, 1, 2)),
    data = ToothGrowth)

Residuals:
   Min     1Q Median     3Q    Max
 -8.20  -2.72  -0.27   2.65   8.27

Coefficients:
```

```
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      20.6633     0.6630  31.166  < 2e-16 ***
suppVC                                           -3.7000     0.9376  -3.946 0.000231 ***
ordered(dose, levels = c(0.5, 1, 2)).L            9.0722     1.1484   7.900 1.43e-10 ***
ordered(dose, levels = c(0.5, 1, 2)).Q           -2.4944     1.1484  -2.172 0.034254 *
suppVC:ordered(dose, levels = c(0.5, 1, 2)).L     3.7689     1.6240   2.321 0.024108 *
suppVC:ordered(dose, levels = c(0.5, 1, 2)).Q     2.7312     1.6240   1.682 0.098394 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.631 on 54 degrees of freedom
Multiple R-Squared: 0.7937,     Adjusted R-squared: 0.7746
F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

This model shows the predictors are significant as categorical predictors as well. Note the difference between the two models: The ancova model fit a different slope and intercept to each delivery method, while the anova model fit a different mean to each dose-delivery method combination. The ancova method uses fewer degrees of freedom, and thus has less flexibility to nonlinear response to dose. The anova model uses more degrees of freedom, however it can capture nonlinear effects by dose.

The question of which model is the right model–well, here is probably does not matter much. In other cases, the question to be answered is whether there is a strong likelihood of nonlinearity in the dose. If there is, the two-way anova model should probably be used. If not, then the ancova model should be used.

Lecture 8

Heteroscedasticity

8.1 Main ideas

- Checking model assumptions
  - Residual vs. fitted plots
  - qq plots
- Box Cox transformations

8.2 `R` Functions

8.3 Testing model assumptions

The ordinary least squares linear model assumes observations are normally distributed $N(X\beta, \sigma^2)$. The assumption of constant variance is critical to inference, in most cases. Non-constant variance (that is, different observations have different variances) is termed heteroscedasticity.[1]

There are two problems: identifying when data might be heteroscedastic and then fixing it in some way. The two methods for identifying heteroscedasticity is quantile-quantile plots (QQ plots) and fitted versus residual plots. This lecture will only cover one method of dealing with data which show heteroscedasticity: power transforms. An alternative, using generalized linear models, will be covered later in the course.

8.3.1 Fitted-residual plot

Residuals have been extensively discussed already–they are the difference between the observations and the predicted observations.[2] That is, the residuals are a vector of the same length as the data, given by $r_i = y_i - \hat{y}_i$.

[1] Pretentious? Perhaps.

[2] Strictly speaking these are "raw" residuals. Often the residuals are standardized by dividing by the estimate of the common variance $\hat{\sigma}^2$, giving the so-called "standardized residuals." Discussion in next lecture will also cover "studentized" residuals.

Plotting the fitted values versus residuals $(\hat{y}, r_i)$ can give some insight into the model adequacy. Residuals should be centered around zero, and if the model holds true, should have no pattern. Common problems include a cone shaped residuals (residuals growing in magnitude as the prediction increases), or apparent nonlinearities. See figure 8.3.1.

8.3.2 QQ plots

Quantile-quantile plots are a common method of examining data normality.[3] To check if a collection of values are normal, the data are first ordered and then plotted against their corresponding quantile from the standard normal distribution. That is, the empirical quantiles of the data are plotted against the theoretical quantiles of the standard normal.[4]

Recall that the ordinary least squares model, the residuals should be normally distributed. Thus plotting the residuals against the normal quantiles should form a straight line. Deviations from the assumption of normality are indicated by deviations from the straight line. See figure 8.3.2.

Note that it is often difficult to determine precisely what the problem is with the data (the differences in QQ plots for non-normal data often look similar), but it does give an indication of a possible problem. Problems which can sometimes be seen on QQ plots include heavy tails or thin tails (that is, more or less data far away from the mean than expected under a normal model), or skew (heavier tail on one side of the distribution).

Formal tests

Although not part of this course, there are a wealth of formal hypothesis tests for normality of data. Common approaches include

- χ^2 test. Group observations into a group, and determine the expected number of observations in those groups under a normal distribution. Although it is used, it fails to reject the null hypothesis when it should more often than the official p value indicates.
- Anderson-Darling test
- Kolmogorov-Smirnov test
- Shapiro-Wilk test
- Cramér-von-Mises test

Each of the tests is sensitive to different types of departure from normality. Both Anderson-Darling and Kolmogorov-Smirnov tests are used to test other hypotheses as well.

8.4 Fixing heteroscedasticity

There are several approaches to "fixing" heteroscedastic data. Perhaps the best approach, however, is to not fix the data but fix the model. Models which use other error distributions are covered

[3] A very similar plot is the Rankit plot–look and interpretation are nearly identical.

[4] Almost. Identifying a particular point with a quantile has a subtly that the end points will either be the 0th or 100th quantile, which have no corresponding value for the normal. Thus commonly (I think this is what is done in `R`), the kth statistic of n total is identified with the $k/(n+1)$ quantile.

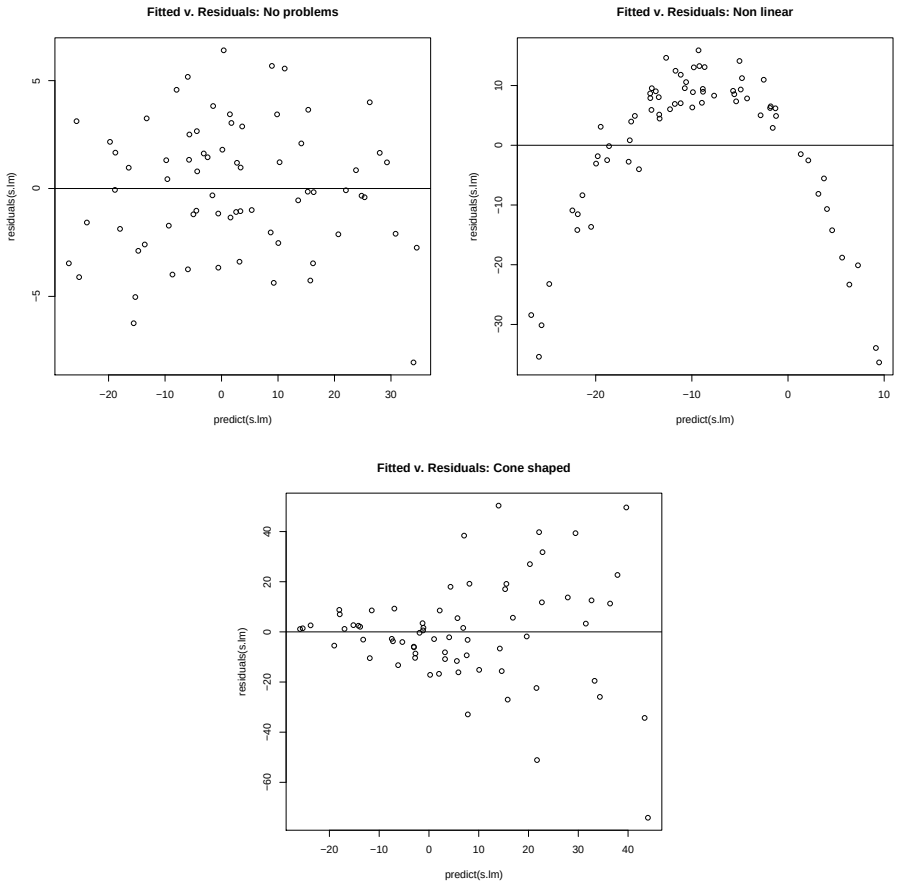


Figure 8.1: Common problems in a fitted v. residuals plot. The first plot (top left) shows no problems. The next plot (top right) shows nonlinearities evident in the residuals, possibly indicating the need for a transformation of one of the predictors. The third plot (bottom) shows a cone shape, which indicates variance which is increasing with the prediction, a common type of heteroscedasticity.

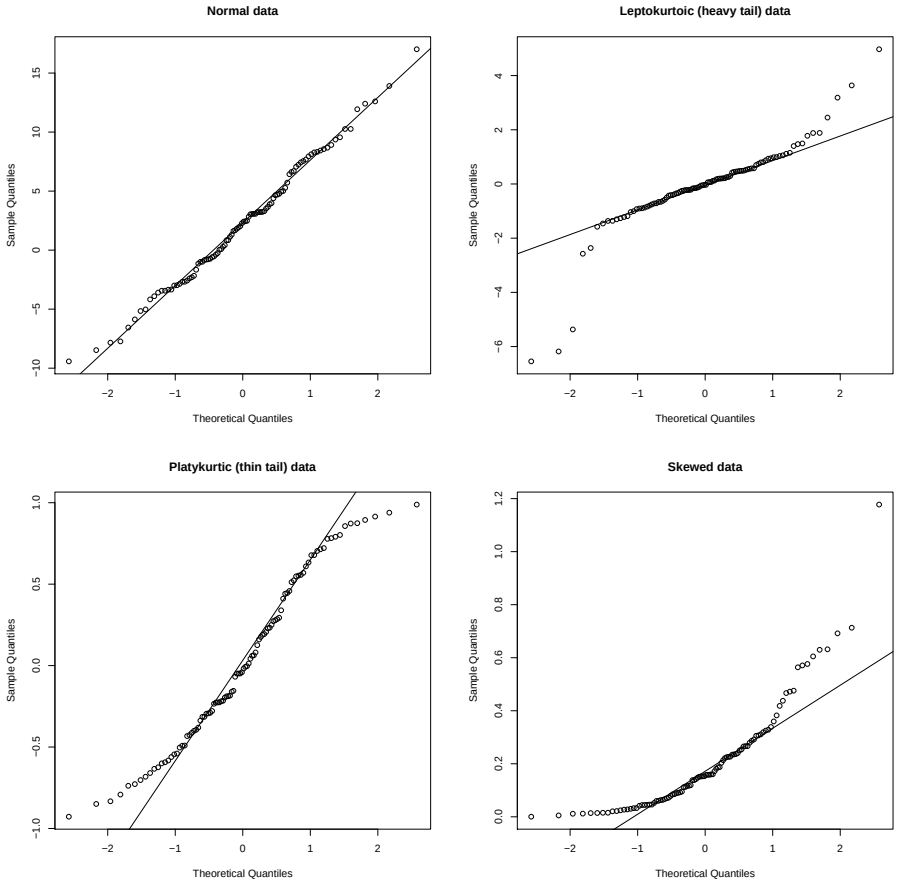


Figure 8.2: QQ plots of 100 generated data from various distributions. Top left, normal data. Top right, students t on 3 degrees of freedom, heavy tails (leptokurtic). Bottom left, uniform data, thin tails (platykurtic). Bottom right, gamma(1,5), which is skewed. Although it is often difficult to precisely identify the problem with the data, a qq plot can indicate when the data are non-normal.

later, as generalized linear models. A more traditional approach is so-called variance stabilizing transformations. That is applying a function to the response which has the effect of making the errors approximately normal.[5]

There are many possible choices for functions to stabilize variance. The most popular (and the only one which will be covered here) is the power transformation, also called the "Box-Cox transformation" after the originators George Box and David Cox.[6] The transformation, which has a free parameter λ, is shown in equation 8.1.

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \tag{8.1}$$

The notation $y^{(\lambda)}$ was introduced by Box and Cox to denote the transformed y. The choice of λ is largely driven by the lambda which best at making the variance approximately equal (more on this below).

The choice of the functional form may seem a bit odd. Keep in mind, for regression purposes, adding a constant or dividing the response by a constant does not have a important impact on the analysis, thus it is rather like doing an analysis on the transformation y^λ.[7] It is important to recognize that this is one of many possible functions which might be used. (Box and Cox even introduced a second function in their paper, and many more have entered the literature since.)

NOTE: For the Box-Cox transformation to make sense, however, all of the values of y must be positive.

8.4.1 Boxcox plots, choosing λ

Finding λ is, in a sense, simply a matter of estimating another model parameter. For estimation of model parameters, one of the most common methods is maximum likelihood. (In some rare cases, it may be possible to make some kind of physical argument as to the correct value of λ. If this is possible, it is probably the best method to go with, but it is difficult to do in practice.) Finding the maximum likelihood estimate is a bit of computation, which has been built into the `boxcox()` function in `R`.[8] By default, it outputs the likelihood function for λ and 95% confidence intervals, see figure 8.4.1.

Since the estimation of an additional parameter could have a significant impact on inference, the exact MLE is not always used. Often a nearby value which has either physical interpretation or an integer or low-order fraction (1/3 or 3/4, for example) is used, since such are likely to have a reasonable physical interpretation. The response is then transformed and the analysis continues as it would for any other linear regression. Is this cheating? Maybe. Is it done? Yes.

[5] For those of you who took Stat 512 and calculated variance stabilizing transformations using the $\int 1/\sigma(\theta)d\theta$ method, that method works well if the true distribution is known. In most cases, however, the true distribution is unknown.

[6] The curious alignment of their names (whose full names are George Edward Pelham Box and David Roxbee Cox) is not coincidental, at least according to some tellings of the story. In fact, the paper was written by these two precisely because their names would correspond to a comic one-act play called Cox and Box.

[7] This is what Box and Cox say, but it is note entirely true. It is true that it does not alter the mechanics of the regression, however the individual values of the β estimates will be changed, and the significance of the β values may be as well. Additionally, the significance may be effected in more subtle ways as well, as is discussed below.

[8] The `boxcox()` function is built into the `MASS` library which will need to be loaded before `boxcox()` can be used.

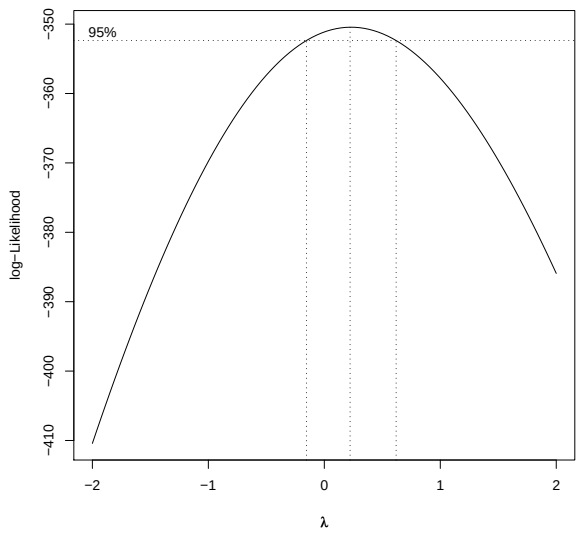


Figure 8.3: An example of the output for a `boxcox()` plot. The log-likelihood is shown as a function of λ. The highest MLE for λ is shown as the peak of this plot. It is common practice, however, to take low-order rational numbers or nearby integers for λ, so as to have a minimal impact on the inference.

8.5 Other important diagnostics (which are skipped in this class)

How many ways can the data fail to be consistent with the model? Quite a lot. Thus there are numerous other diagnostics which are commonly used to determine if other problems crop up. Two common problems which have not been discussed here are strongly correlated predictors, and correlated errors. Correlated predictors can lead to problems for the fit parameters $\hat{\beta}$ become highly correlated. Correlated errors, on the other hand, is a violation of the independence assumption, and substantial effort has gone into dealing with them (as time series or longitudinal data) using different models.

Highly correlated predictors mean most of the information is about some linear combination of β, and not β itself. For instance, if the true values of β_1 and β_2 are 2 and 3, and $x_1 \approx x_2$, it is likely that $\hat{\beta}_1 = 3$ and $\hat{\beta}_2 = 2$ would fit nearly as well. In fact $\hat{\beta}_1 = 0$ and $\hat{\beta}_2 = 5$ would probably also work pretty well. Here, we really only have information about $\beta_1 + \beta_2$ in the sense that estimates which satisfy $\hat{\beta}_1 + \hat{\beta}_2 = 5$ will likely be almost as good as the true values of 2 and 3.[9]

Correlated errors can mean any number of things. Linear models are occasionally used for time series, however this is often inappropriate. Correlated residuals is an indicator of this problem, but they might arise in other circumstances as well. If the data clearly have a nonlinear response to the predictor (exhibiting, for example, a curved response), this generally shows up as trends in the residuals (see figure 8.3.1). There are formal tests, such as the runs-test, for determining if the residuals are not really randomly distributed in some choice of sequence (usually the order the data were collected).

8.6 Examples

8.6.1 Trees

In harvesting trees there are many variables of interest, however the central variable of interest is volume. It is difficult, however, to get a good sense of tree volume without cutting the tree. Diameter[10] and height are quite easily measured, on the other hand, and it might be reasonable to suspect that volume is closely related to both.

To look for a relationship, a linear model can be applied. Data on diameter, height and volume from 31 trees is described in the table below.[11]

[9] In fact, if $x_1 = x_2$, the estimates for β_1 and β_2 would be *identical* so long as $\hat{\beta}_1 + \hat{\beta}_2 = 5$, which is why $(X'X)$ would not be invertible.

[10] Often tree "diameter" of a tree trunk is measured in DBH: diameter at breast height. What constitutes "breast height" varies but is generally 1.3 to 1.5 meters off the ground according to Wikipedia.

[11] These data are from Loveday's notes on the class; I don't know their original source, but they appear on numerous other course websites and seem to be a common dataset for this sort of example.

Diameter	Height	Volume	Diameter	Height	Volume
8.6	65	10.3	12.9	85	33.8
8.8	63	10.2	13.3	86	27.4
10.5	72	16.4	13.7	71	25.7
10.7	81	18.8	13.8	64	24.9
10.8	8	19.7	14	78	34.5
11	66	15.6	14.2	80	31.7
11	75	18.2	14.5	74	36.3
11.1	80	22.6	16	72	38.3
11.2	75	19.9	16.3	77	42.6
11.3	79	24.2	17.3	81	55.4
11.4	76	21	17.5	82	55.7
11.4	76	21.4	17.9	80	58.3
11.7	69	21.3	18	80	51.5
12	75	19.1	18	80	51
			20.6	87	77

We might guess that the relationship between these variables is not immediately linear. Specifically, since trees to a first order of approximation are cylindrical,

$$v \approx \frac{1}{4}\pi d^2 h, \tag{8.2}$$

where V is volume, d is diameter, and h is height. Transforming this equation using a log (no pun intended), yields

$$\log v = \log\left(\frac{1}{4}\pi\right) + 2\log(d) + \log(h). \tag{8.3}$$

This is a linear equation. The best course of action here is probably just to fit the model $\log(y) = \beta_0 + \beta_1 \log(d) + \beta_2 \log(h)$. However for illustration purposes, let's try a boxcox plot to find the best transformation of y.

First, read in the data and log transform the predictors

```
> trees<-read.table(
      file='http://students.washington.edu/nesse/qerm514/data/treesboxcox.txt',
      header=T)
> dm<-log(trees$diam)
> vm<-trees$vol
> ht<-log(trees$height)
```

To generate a boxcox plot, use the `MASS` library and the `boxcox()` command.

```
> library(MASS)
> boxcox(lm(vm~dm+ht))
```

The result is shown in figure 8.4.

The boxcox plot shows a maximum for λ very near 0. Consulting equation 8.1, this suggests a log transformation of the response (volume) is indeed appropriate.

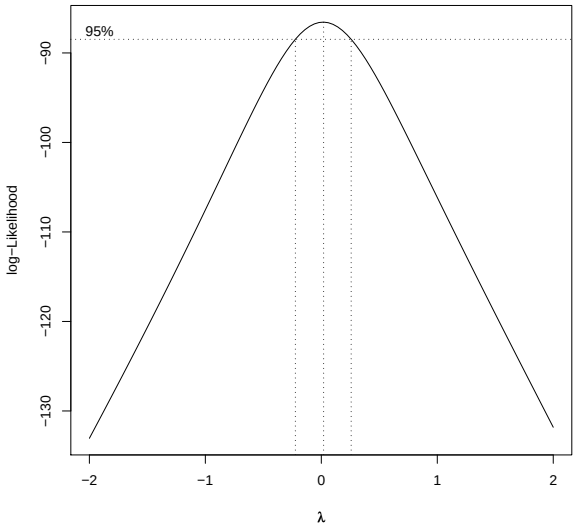


Figure 8.4: The results of the boxcox plot of the trees data (see section 8.6.1). The data show a log transformation of the response is likely appropriate.

8.6.2 Beer

Although the contents of these lectures is primarily focused on ecology, important results was recently published in physics which may be important to work here. The results were published in 2002 in the *European Journal of Physics* (vol 23: 21-26) in a report titled "Demonstration of the exponential decay law using beer froth." The author, Arnd Leike, poured several beers and measured the height of the head of foam[12] every few seconds for six minutes. He then fit these data to an exponential decay model, which he argues on theoretical grounds is the appropriate model.

The model fit in the paper relates height of the foam h to time t with a free parameter τ. The initial height of the beer h_0 is taken to be the first observation.

$$h = h_0 e^{-t/\tau} \tag{8.4}$$

Leike tests three beers Erdinger Weissbier, Augustienerbräu Müchen, and Budweiser Budvar (not the American company). The data are shown below. (The units which are listed are cm^{-1}. As near as I can tell, this is a strange way of saying the measurements are in centimeters–thus multiplying by cm^{-1} results in unit-less values displayed. It is rather odd.)

[12]This brings up the critical question, is beer foam good or bad? In some circles, techniques are employed to minimize the head on a beer pour; others, however, argue that a head of foam minimizes carbon dioxide loss and thus is a good thing. There are even beer glasses specifically designed with bubble nucleation sites so as to ensure foam occurs.

time	weiss	munchen	budvar
0	17	14	14
15	16.1	11.8	12.1
30	14.9	10.5	10.9
45	14	9.3	10
60	13.2	8.5	9.3
75	12.5	7.7	8.6
90	11.9	7.1	8
105	11.2	6.5	7.5
120	10.7	6	7
150	9.7	5.3	6.2
180	8.9	4.4	5.5
210	8.3	3.5	4.5
240	7.5	2.9	3.5
300	6.3	1.3	2
360	5.2	0.7	0.9

Observation i is modeled as

$$h_i = h_0 e^{-t_i/\tau} + \epsilon_i. \tag{8.5}$$

Leike fit this model using techniques which have not yet been covered in the course.[13] His fits for τ however, are 276, 124, and 168 for the Weissbier, Müchen and Budvar respectively. Using these data, it is possible to examine the assumption of normal distribution of the observations.

The raw residuals of the data, under Leike's assumption, should be normal. To examine this assumption, the fit curve, the residuals versus fitted, and the qq plot is shown for the Weissbier, see figure 8.6.2.

```
> ## Import the data
> beer <- read.table(
      file='http://students.washington.edu/nesse/qerm514/data/beer.txt',
      header=T)
> ## Plot the data
> plot(beer$time,beer$weiss)
> lines(beer$time,beer$weiss[1]*exp(-beer$time/276))
> ## Fitted-residuals plot
> res<-beer$weiss[2:15] - beer$weiss[1]*exp(-beer$time[2:15]/276)
> plot(beer$weiss[1]*exp(-beer$time[2:15]/290),res,main="Weissbier fitted-resid")
> abline(h=0)
> ## QQ-plot
> qqnorm(res)
> qqline(res)
```

The data show remarkable departure from the normality assumption. There is a fairly clear trend in the residuals, as shown in the residuals versus fitted plot, and likely non-normal tails as shown in the qq plot. This is not, perhaps, unsurprising. It is probably reasonable to suspect heteroscedasticity just from the setup of the problem, since as the froth height goes to zero the magnitude of the likely observational error is likely to get smaller (it is much easier to mistake 17.5

[13] Note that this model is accessible to the techniques we've used so far via a log transformation of the observations. That assumes a multiplicative log-normal error, however, whereas Leike used an additive normal error. For reasons which are somewhat mysterious to me, he took repeated measurements but only fit the average, thus losing lots of degrees of freedom. In several ways the paper has problems–fixing some of these problems is an extra credit problem.

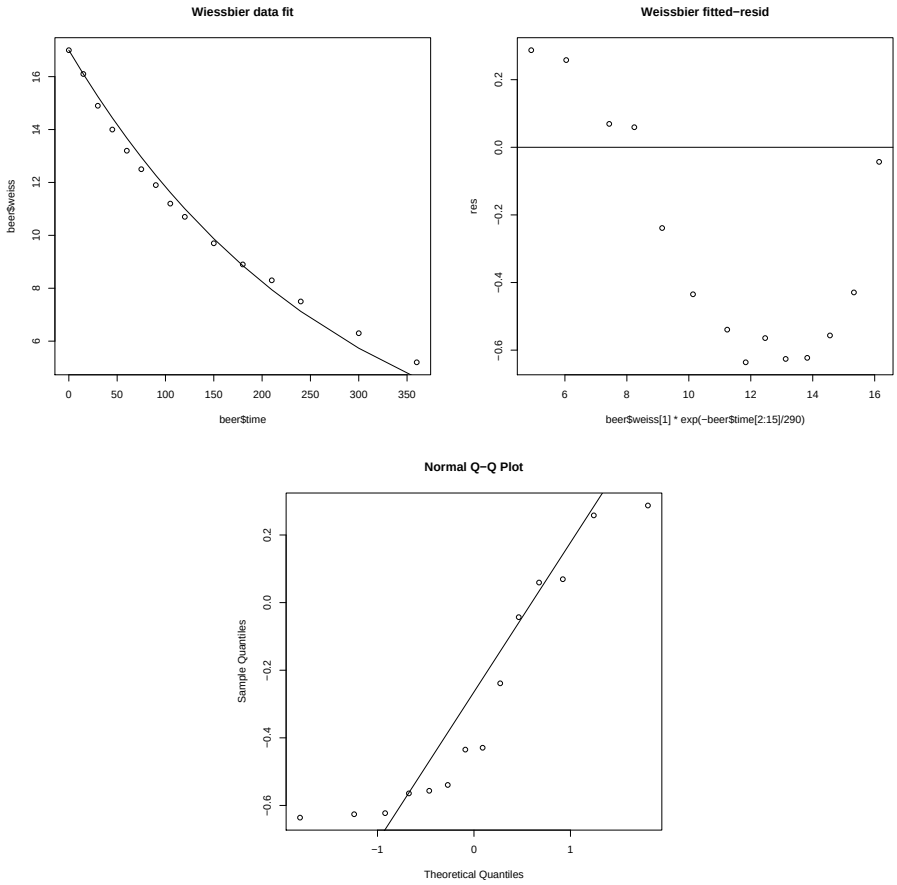


Figure 8.5: The data, fitted-residuals, and qq plot for the Weissbier measurements from the beer froth paper. Note that the qq plot and residuals versus fitted seem to show systematic departure from normality.

for 17 than mistake 0.1 for 0.6). The trend in the residuals, however, suggest more than a poor choice of model for errors–it suggests the exponential model may be the wrong one here.

Notably, the other two beers do have better fits to the normality assumption (as judged from the qq-plot), they all show a similar trend in the fitted-residuals plot. This is strongly suggestive that the main point of the paper[14] is incorrect–beer froth does not have an exponential decay.

[14] well, other than education about the fitting technique

Lecture 9

Outliers and extreme observations

9.1 Main ideas

- Hat matrix
- Finding "outliers"
  - Leverage
  - D-fits
  - Cook's distance
  - Studentized residuals
- Corrective measures

9.2 Hat matrix

Rank and trace

Solving the least squares problem $Y = X\beta + \epsilon$ yielded $\hat{\beta} = (X'X)^{-1}X'y$. Thus predicted observations $\hat{y} = X(X'X)^{-1}X'y$. The first part of this expression $H = X(X'X)^{-1}X'$ is usually called the "hat matrix." The hat matrix is square and $HH = H$ (a property which is called idempotent).

Theorem 9.2.1 (Hat matrix)**.** *The rank of the hat matrix is p, the number of parameters in the model (this was actually used in theorem 4.4.1 to apply Cochrane's theorem). Furthermore, the sum of diagonal elements (usually called the trace) of the hat matrix is also p.*

Proof. (Rather "proof" in a sketchy way) To see the rank, think of this geometrically. The expected value of y, $\mathrm{E}Y = X\beta$ must exist in a p dimensional subspace of $\mathbb{R}^n$ (to see this, note that there are p columns, x_1 through x_p of X, and $X\beta$, as β varies, the span of x_1 though x_p). The hat matrix is precisely the matrix which projects onto this subspace, and thus must have rank p.

The proof of the trace-rank equality is a direct consequence of the rank. Using the definition of eigenvalues (and the fact H is idempotent), $Hv = H^2v = \lambda x = \lambda^2 x$, implying $\lambda^2 = \lambda$. Thus

λ must equal either zero or one. Using the fact (unproved here) that the rank of a matrix is the number of non-zero eigenvalues, and that the trace is the sum of eigenvalues, the rank must equal the trace. □

Relation to residuals

It initially seem a bit peculiar, but the (raw) residuals only estimates of the true errors on each observation. In the model $y = X\beta + \epsilon$, we can think of the residuals $y - \hat{y}$ as being estimates of ϵ, however the true values of ϵ are unknown. In this context, it is common to describe the estimated error $\hat{\epsilon}_i = y_i - \hat{y}_i$ on a particular observation i.

Definition 9.2.1 (Leverage)**.** *The leverage h_{ii} of an observation i is the ith diagonal element of the hat matrix H.*

The variance we could expect for an observation which has high leverage should be smaller than the variance for an observation which has low leverage, since high leverage forces $\hat{y}$ to be close to y. In fact, there is a theorem relating the leverage to the variance of the error estimate $\hat{\epsilon}$.

Theorem 9.2.2 (Error estimates)**.** *The variance of the estimated error $\hat{\epsilon}_i = y_i - \hat{y}_i$ is equal to $\sigma^2(1 - h_{ii})$, where h_{ii} is the leverage for observation i.*

Proof. The variance of $\hat{\epsilon}_i = y_i - \hat{y}_i$ can be broken down (in general) as $\text{Var}\hat{\epsilon} = \text{Var}Y + \text{Var}\hat{Y} - 2\text{Cov}(Y, HY)$. A bit of matrix calculation shows this is $\sigma^2(I + H - 2H)$, which is just $\sigma^2(1 - h_{ii})$ for the ith observation. □

9.3 Finding unusual observations

The term "outlier" sometimes gets a formal definition, so it may be best to use "unusual observation." An unusual observation is only unusual in the context of a model, and for the usual model $Y \sim N_n(X\beta, \sigma^2 I)$, an unusual observation is one which is far away from its predicted value (far being relative to the estimated σ^2). Formally, an unusual observation is a probable violation of the model assumptions. Sometimes it is evident just from looking at the data, however more often than not some other techniques are needed.

A related problem is the detection of influential observations–those which greatly influence the regression even if they are close to the prediction. Although highly influential points do not violate model assumptions, a regression which is highly dependent on the values of one or two data should be looked at with a skeptical eye.

The sole goal of this lecture is detection. What to do about usual observations is a substantial topic unto itself and won't be covered here (a brief overview of methods is included, but only enough to be dangerous).

9.3.1 Leverage

The diagonal elements h_{ii} of the hat matrix H give a bit of information about the relative importance of the observation in determining the parameters. Note the leverage is entirely dependent on X and there is a value of leverage for each observation. By theorem 9.2.1, the sum of the n leverages is p.

Without putting a distribution on X (which has been thus far avoided), it is impossible to give a theoretical distribution of leverage values h_{ii}. A rule-of-thumb for a large leverage, however, is

$$h_{ii} > 2\frac{p}{n}. \tag{9.1}$$

Since the sum of the leverage values is p and there are n of them, it is reasonable to expect that most of the leverage values would be $\approx p/n$. High leverage, of course, does not mean the point is an outlier, but it may be worth looking more closely at since it has a large influence on the predictions.

9.3.2 Studentized residuals

Using theorem 9.2.2 the estimated standard deviation of an error estimate $\hat{\epsilon}$ is $\hat{\sigma}\sqrt{1-h_{ii}}$. A quantity based on this property is the studentized residual.

Definition 9.3.1 (Studentized residual)**.** *The studentized residual r_i for observation i is defined as*

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{9.2}$$

Studentized residuals are a way of examining which observations are large, relative to their leverage. This is the third form of residuals (the other two being raw and standardized). A variation on this is to use the estimated standard deviation of the residuals, not including the ith point. Such a residual is commonly called "externally studentized" and is denoted with a t_i.

9.3.3 Dffits

Perhaps the most intuitive way of assessing the impact a single point has on a model's fit is to fit the model twice–once with all the data, and once with the point removed. The general term for this technique is "cross-validation" (and it will come up again in model selection), however for now denote $\hat{y}_{[i]}$ as the vector of estimated y, fit to data with the ith point removed. Thus $Y_{i(i)}$ is the ith predicted observation which was fit to the data missing the ith observation.

The long way to calculate dffits is

1. Calculate the usual model fit $\hat{y} = X(X'X)^{-1}X'y$.
2. Remove the ith row from X (call this $X_{[-i,]}$) and the ith element from y (call this $y_{[-i]}$).
3. Fit the model as usual, $\hat{y}_{[-i]} = X_{[-i,]}(X'_{[-i,]}X_{[-i,]})^{-1}X'_{[-i,]}y_{[-i]}$ and $\hat{\sigma}^2_{[i]} = \frac{1}{n-p-1}(y_{[-i]} - \hat{y}_{[i]})'(y_{[-i]} - \hat{y}_{[i]})$.
4. Define the DIFFITS for each observation $DFFITS_i = (\hat{y}_i - \hat{y}_{[-i]})/\sqrt{\hat{\sigma}^2_{[-i]}h_{ii}}$ (where h_{ii} is the leverage).

Note that this method, however, would require calculating a new linear regression for each data point. This is computationally quite intensive. An equivalent definition (see theorem 9.3.1) is much simpler to calculate.

$$DFFIT_i = t_i\sqrt{\frac{h_{ii}}{1-h_{ii}}} = \frac{\hat{y}_i - \hat{y}_{[-i]i}}{\hat{\sigma}_{[-i]}\sqrt{h_{ii}}} \tag{9.3}$$

where t_i is the externally studentized residual, see equation 9.2. Note that there is a DFFITS for each observation i.

Theorem 9.3.1 (DFFITS definition equivalence)**.** *The two definitions of DFFITS, given in equation 9.3 and step 4 above, are equivalent. That is*

$$DFFIT_i = (\hat{y}_i - \hat{y}_{[i]})/\sqrt{\hat{\sigma}^2_{[i]} h_{ii}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}. \tag{9.4}$$

Proof. Proof is given in Appendix C of Montgomery and Peck's 1992 book *Introduction to linear regression analysis*. Sketched, the proof relies on first showing the inverse of the the product of the i, ith minor of X, $X_{[-i,-i]}$ with itself is

$$(X'_{[-i,-i]} X_{[-i,-i]})^{-1} = X'X + \frac{(X'X)^{-1} X_i X_i' (X'X)^{-1}}{1 - X_i'(X'X)^{-1} X_i}. \tag{9.5}$$

With a good deal of (non-trivial) calculation, this can be used to show

$$\hat{\beta}_i - \hat{\beta}_{(i)} = \frac{(X'X)^{-1} X_i (y_i - \hat{y}_i)}{1 - h_{ii}} \tag{9.6}$$

This is rearranged to give the result. □

9.3.4 Cook's distance

Dffits have the drawback that they only measure the effect removal has on one point. Another measure, Cook's distance, indicates the influence removal of a point has on all of the points.

Definition 9.3.2 (Cook's distance)**.** *Cook's distance D_i describes an aggregate effect of the removal of observation i.*

$$D_i = \frac{(\hat{y} - \hat{y}_{[i]})'(\hat{y} - \hat{y}_{[i]})}{p\hat{\sigma}^2} \tag{9.7}$$

The distribution of D_i under the null hypothesis is approximately $F_{p,n-p}$.[1] Using this, points with a Cook's distance which is greater than 1 might be worth looking at carefully; most of the observations should be much smaller than this, however, so perhaps a stricter standard is needed. Ultimately, of course, Cook's distance is just a means of indicating *possible* issues, so there is no strict hypothesis test or similar algorithm here.

9.3.5 Summary of case statistics

There are a lot of methods described for identifying residuals. The above methods are generally case statistics–values which are dependent on the case i. Below is a table of the case statisitcs which may be useful in identifying outliers or influential observations.

[1] As far as I can tell, this is not a rigorous result. Two issues arise as far as I can see: it is not obvious to me that the numerator is independent of the denominator, and what is really under consideration is the maximum D_i not a specific D_i, when the question is what constitutes an unusual observation. Further exploration is an extra credit assignment.

	Symb.	Formula	Description
Raw residual	$\hat{\epsilon}_i$	$y_i - \hat{y}_i$	The difference between the fitted value and the observation. Note that it may be fairly small if the point is highly influential, even if it came from a different underlying distribution.
Standardized residual		$\frac{y_i - \hat{y}_i}{\hat{\sigma}}$	The raw residual divided by the common estimate of the standard deviation. Similar to raw residuals in interpretation.
Leverage	h_{ii}	$(X(X'X)^{-1}X')_{ii}$	Gives an indication of the influence of a point, without relying on y. Does not indicate if the point is an outlier, only if it is influential.
Internally Studentized residual	r_i	$\frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$	Combines influence and how far the point is off the line.
Deleted residual	d_i	$y_i - \hat{y}_{[-i]i}$	The distance from the point to the line fit without the point.
Externally studentized residual	t_i	$\frac{y_i - \hat{y}_{[-i]i}}{\hat{\sigma}_{[-i]}(1-h_{ii})^{-1}}$	A Studentized version of the deleted residual.
DFFITS	$DFFITS_i$	$\frac{\hat{y}_i - \hat{y}_{[-i]i}}{\sigma_{[-i]}\sqrt{h_{ii}}}$	A standardized measure of the fit of the point i with and without point i included.
Cook's Distance	D_i	$\frac{(\hat{y}-\hat{y}_{[-i]})'(\hat{y}-\hat{y}_{[-i]})}{p\hat{\sigma}^2}$	A measure of how well the model changes at all points by removing point i.

9.4 Remedial measures

9.4.1 Throwing out points

The most primitive, perhaps, strategy is to throw out the unusual observations, particularly if the point dramatically changes the fit. This is sometimes justified (such as values which so strange that they almost certainly resulted from a recording error). It is probably not justified as often as it is employed, however. The danger is that the data represent real observations; throwing them out

may be throwing out valuable information. More subtly, selectively throwing out points introduces the possibility of arriving at misleading inference (if, for instance, a researcher throws out the data which disagree with their conclusion).

If there are possible, but not identifiable problems with data, it may be reasonable to do (and report) the analysis twice–once with the observation and once without. In a sense, this is carrying the uncertainty in the observation's validity forward into uncertainty in the conclusion.)

9.4.2 Robust regression

Recall that least squares is a procedure for fitting. Although it has nice properties in the context of the classical linear model, it is not the only option available for fitting. At the expense of some nice properties, the influence of extreme observations can be reduced by choosing another fitting routine. Commonly, this amounts to choosing another function $\rho(y - \hat{y})$. Recall least squares minimizes

$$(y - \hat{y})'(y - \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{9.8}$$

Generalizing this, we could instead minimize

$$\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9.9}$$

or more generally for some function ρ,

$$\sum_{i=1}^{n} \rho(y_i - \hat{y}_i). \tag{9.10}$$

There are several other approaches discussed in Faraway's *Linear models with* `R`.

9.5 Examples

Brain-body	9.5.1	Identifying potential outliers in a linear regression.
Pot busts	9.5.2	Identifying outliers in multiple regression.

9.5.1 Brain body weight

Brain size has long held fascination for those seeking to separate humans from the other animals.[2] Humans certainly do not have the largest brains,[3] but many people have suggested that perhaps a high brain-body ratio is more important.

The data below are 62 brain and body weights from `R`'s built-in library `mammals`.

[2]Measuring brain size in humans to get at intelligence was also popular among 19th century anthropologists, statisticians, and anatomists. A fascinating history of this largely racist endeavor (and the flaws therein) can be found in Stephen J. Gould's *The Mismeasure of Man.*

[3]I found conflicting information indicating elephants or whales were largest, but didn't really want to sort it out.

	Body	Brain		Body	Brain
Artic fox	3.385	44.50	Human	62.000	1320.00
Owl monkey	0.480	15.50	African elephant	6654.000	5712.00
Mountian beaver	1.350	8.10	Water opossum	3.500	3.90
Cow	465.000	423.00	Rhesus monkey	6.800	179.00
Grey wolf	36.330	119.50	Kangaroo	35.000	56.00
Goat	27.660	115.00	Yellow-bellied marmot	4.050	17.00
Roe deer	14.830	98.20	Golden hamster	0.120	1.00
Guinea pig	1.040	5.50	Mouse	0.023	0.40
Verbet	4.190	58.00	Little brown bat	0.010	0.25
Chinchilla	0.425	6.40	Slow loris	1.400	12.50
Ground squirrel	0.101	4.00	Okapi	250.000	490.00
Artic ground squirrel	0.920	5.70	Rabbit	2.500	12.10
African giant pouched rat	1.000	6.60	Sheep	55.500	175.00
Lesser short-tailed shrew	0.005	0.14	Jaguar	100.000	157.00
Star-nosed mole	0.060	1.00	Chimpanzee	52.160	440.00
Nine-banded armadillo	3.500	10.80	Baboon	10.550	179.50
Tree hyrax	2.000	12.30	Desert hedgehog	0.550	2.40
N.A. opossum	1.700	6.30	Giant armadillo	60.000	81.00
Asian elephant	2547.000	4603.00	Rock hyrax-b	3.600	21.00
Big brown bat	0.023	0.30	Raccoon	4.288	39.20
Donkey	187.100	419.00	Rat	0.280	1.90
Horse	521.000	655.00	E. American mole	0.075	1.20
European hedgehog	0.785	3.50	Mole rat	0.122	3.00
Patas monkey	10.000	115.00	Musk shrew	0.048	0.33
Cat	3.300	25.60	Pig	192.000	180.00
Galago	0.200	5.00	Echidna	3.000	25.00
Genet	1.410	17.50	Brazilian tapir	160.000	169.00
Giraffe	529.000	680.00	Tenrec	0.900	2.60
Gorilla	207.000	406.00	Phalanger	1.620	11.40
Grey seal	85.000	325.00	Tree shrew	0.104	2.50
Rock hyrax-a	0.750	12.30	Red fox	4.235	50.40

Note first that humans do not have the largest brain to body ratio–fully 8 species have a higher ratio (with the largest being the ground squirrel). A plot of the data shows a lack of linearity (see figure 9.1). However a log transformation does show some linearity.

```
> plot(mammals$body,mammals$brain,
    main="Mammal brain sizes",
    xlab="Body weight (Kilograms)",
    ylab="Brain weight (grams)")
> plot(log(mammals$body),log(mammals$brain),
    main="Mammal brain sizes",
    xlab="Log Body weight (Kilograms)",
    ylab="Log Brain weight (grams)")
> text(log(62),log(1320)+.2,"Human")
```

The data do not, at least immediately, appear to have any outliers. Humans (the 32nd point in the dataset) are on the high side of the curve, but do not seem to be particularly unusual. To examine how unusual the observation is, the first place to look is perhaps the residuals. Using the `which.max()` command, it is possible to determine which of a collection is the maximum. Below, the maximum residual seems to come from humans.

```
> mam.lm<-lm(log(brain)~log(body),data=mammals)
> which.max(residuals(mam.lm))
```

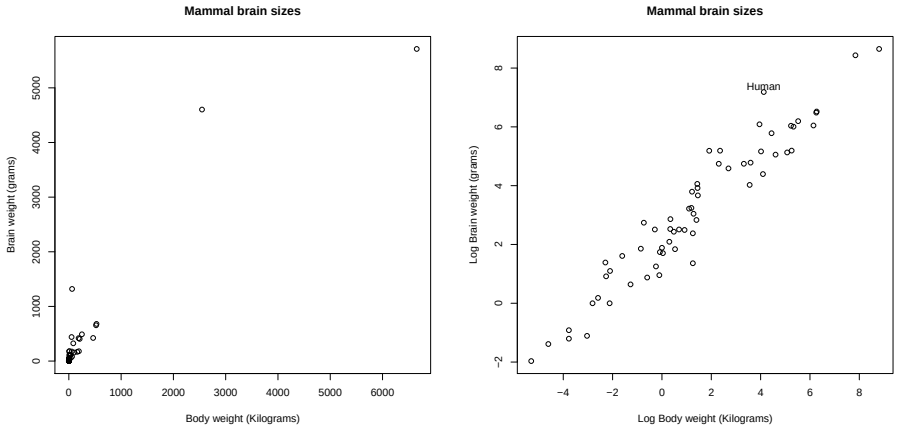


Figure 9.1: Mammal body weights and brain sizes, both untransformed (left) and log transformed (right).

```
Human
   32
```

Recall that residuals are useful for finding unusual observations if the points themselves are not overly influential. Here the influence is fairly small, compared against the usual rule of $\frac{2p}{n} \approx 0.0645$. If the point were influential or this is a concern, Cook's distance is probably a better measure. In fact, Humans also have the highest Cook's distance.

```
> which.max(cooks.distance(mam.lm))
Human
   32
```

There is a more subtle philosophical/scientific question, however: simply because humans are unusual in this way, does that mean culture and intelligence can be attributed to this effect? If so, what is the causal mechanism?

9.5.2 Pot busts

Journalists routinely report on large police drug busts, often reporting the amount and "street value" of the drugs confiscated. In the case of marijuana, the amount is often reported in either plants seized, or pounds seized. The US drug enforcement agency uses the estimating method of $1,000 per pound of marijuana ($62.50 per oz.). This conversion to estimate street value, however, is not universally applied.

To determine the factors which are used to determine the street value reported, a 24 recent recent news stories were culled from Google News[4] which reported both amount of pot seized and some estimate of value.

[4]The most recent reports on the news, on March 7, 2008, using the key words "pot bust."

Weight (lbs)	Plants (cnt)	Value estimated (USD)
56	0	282000
20	152	430000
0	117	117000
0	125	175000
8	0	57000
6700	0	60000000
0	176	176000
265	0	250000
100	0	85000
25	0	125000
25	28	84000
125	0	300000
0	90	144000
60	0	120000
0	96	153000
0	1014	1200000
1229	0	2000000
0	2300	2750000
1916	0	1500000
40	0	40000
15	0	45000
34	0	65000
156.2	0	1400000
126	0	500000

First, read in the data and set up a linear model. Since, with reasonable confidence, the value should go through (0,0,0), the linear model should reflect this.

```
> pot<-read.table(
      file='http://students.washington.edu/nesse/qerm514/data/potbusts.txt',
      header=T)
>
> pot.lm<-lm(pot$val ~ pot$wt + pot$plts+0)
> summary(pot.lm)

Call:
lm(formula = pot$val ~ pot$wt + pot$plts + 0)

Residuals:
      Min        1Q    Median        3Q       Max
-14057329   -406156    -77394      7038   5598066

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
pot$wt     8119.7      519.3  15.637 2.12e-13 ***
pot$plts   1195.3     1452.3   0.823    0.419
```

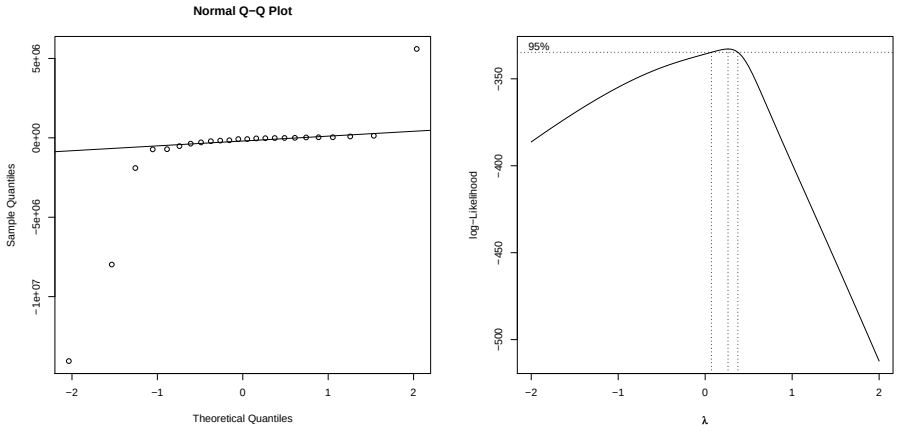


Figure 9.2: The qq and boxplot for the pot value data. The QQ plot seems to indicate heavy tails, while the Box Cox plot seems to indicate use of a transformation of about $\frac{1}{4}$ could be used.

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3680000 on 22 degrees of freedom
Multiple R-Squared: 0.9177,     Adjusted R-squared: 0.9102
F-statistic: 122.6 on 2 and 22 DF,  p-value: 1.179e-12
```

The weight coefficient does strongly indicate that the $1,000 per pound number the DEA recommends is far lower than the media has been using. The plant regression coefficient seems much lower, but may be not significant.[5]

Is the normal model justified? The qq-plot seems to indicate heavy tails, while a Box-Cox plot seems to indicate a transformation of about $\frac{1}{4}$ may be warranted.

```
> qqnorm(resid(pot.lm))
> qqline(resid(pot.lm))
> boxcox(pot.lm)
```

The results are plotted in figure 9.2.

Before making these transformations, however, it may be useful to look at the degree to which this is dependent on outliers or highly influential points. The leverage of the points is shown below.

```
> influence(pot.lm)$hat
          1           2           3           4           5           6
6.244642e-05 3.606822e-03 2.132346e-03 2.433918e-03 1.274417e-06 8.938839e-01
          7           8           9          10          11          12
4.825155e-03 1.398374e-03 1.991276e-04 1.244548e-05 1.345535e-04 3.111369e-04
```

[5]Significant here is a questionable concept since we *know* that plant count is being used in the calculation of seizure value.

```
          13           14           15           16           17           18
1.261743e-03 7.168594e-05 1.435583e-03 1.601629e-01 3.007705e-02 8.240274e-01
          19           20           21           22           23           24
7.310086e-02 3.186042e-05 4.480371e-06 2.301915e-05 4.858403e-04 3.161350e-04
```

The usual rule of thumb on leverage is to identify those points with leverage above $2p/n$, which here is about 0.167. Note that point 6 and point 18 have high leverage according to this rule, and point 16 is on the edge.

Another way to find extreme observations is to use Cook's distance.

```
> cooks.distance(pot.lm)
           1            2            3            4            5            6
6.878359e-08 9.904023e-07 4.127883e-08 5.914041e-08 2.979848e-12 9.185847e+01
           7            8            9           10           11           12
2.125111e-07 1.872646e-04 3.887417e-06 2.795441e-09 1.155179e-07 5.876405e-06
          13           14           15           16           17           18
6.197048e-08 3.569295e-07 7.779162e-08 1.210393e-06 7.516114e-02 7.168003e-07
          19           20           21           22           23           24
6.208516e-01 9.542148e-08 9.756918e-10 3.786910e-08 3.114866e-07 3.196024e-06
```

Again, the points which stand out are 6 and 19. Reading these tables is a bit hard on the eyes, so it helps to graph the data. The `plot()` command, when it gets a linear model object, plots four diagnostic plots of the data, reproduced in figure **??**.

```
> par(mfrow=c(2,2))  #Creates a 2 by 2 grid of plots
> plot(pot.lm)
> par(mfrow=c(1,1))  #Resets the grid to 1
```

Looking at the magnitude of these predictors gives some indication as to why they have high leverage: Point 6 has the largest seizure (by a fair amount) in terms of pounds, while points 16 and 18 are the first and second highest seizure in terms of plants. Point 6, in particular, is pretty substantially away from the rest of the data (although keep in mind a plot like the one shown does not plot both predictors, so it may be misleading in a sense).

```
> plot(pot$wt,pot$val,main="Drug bust value",type='n')
> for(tt in 1:24){
      text(pot$wt[tt],pot$val[tt],as.character(tt))
  }
```

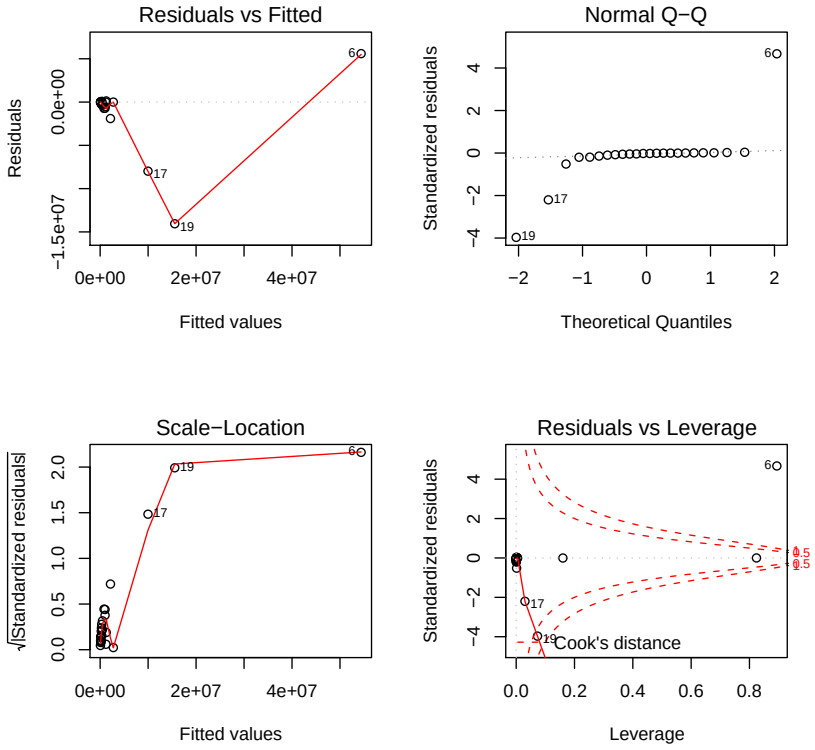
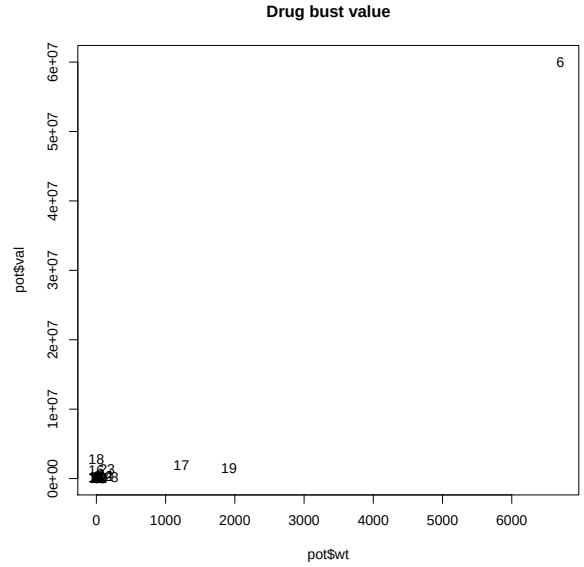


Figure 9.3: Diagnostic plots for the pot bust data. The fitted-residuals plot (upper left) and qq plot (upper right) are described in the previous lecture. The fitted versus standardized residuals is similar to the fitted v. residual plot, only each residual is divided by the common estimate of the standard deviation σ. The square root of the absolute value is then taken. The standardized residuals -leverage plot (lower right) also shows Cook's distance. Note that Cook's distance is a function of the residual and leverage, so the isoclines trace out the Cook's distance for any point in a region.



Although the influential points are perhaps these are good data, it is at least worthwhile looking at the effect on the results to remove the points. If the analysis is too dependent on them, the results might be problematic to interpret. The fit without point 6 is shown below.

```
> pot.r.lm<-lm(pot$val[-6] ~ pot$wt[-6] + pot$plts[-6]+0)
> summary(pot.r.lm)

Call:
lm(formula = pot$val[-6] ~ pot$wt[-6] + pot$plts[-6] + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-572094  -12375   28778  131390 1231075

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
pot$wt[-6]     1081.5      151.7   7.127 4.99e-07 ***
pot$plts[-6]   1199.4      138.3   8.675 2.19e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 350300 on 21 degrees of freedom
Multiple R-Squared: 0.8573,     Adjusted R-squared: 0.8437
F-statistic: 63.06 on 2 and 21 DF,  p-value: 1.326e-09
```

Note that the weight coefficient has dramatically changed. This does seem to indicate that the regression is highly dependent on the single observation.

The course of action for such an influential point depends on the researcher. There is no reason to think that there is anything wrong with the point, so it seems rather unjustified throwing it out. However it is so dramatically influential that, if it is kept, the results are heavily dependent on that one point. Do an analysis with and without the point included, and discuss the results more extensively in the discussion section. While perhaps the most intellectually honest approach, it does not translate well to a non-technical audience (imagine describing this result to a newspaper reporter).

Lecture 10

Linear practice

This is a chapter which has no new concepts, only examples of using linear modeling to estimate effects and interpret those results. The examples are intended to draw on many of the ideas of the past chapters, looking at how linear models can address many a problem.

10.1 Auto theft

Car theft in Seattle, along with numerous other crimes, is recorded every month by the Seattle Police department. Car theft varies between a few hundred to just over a thousand reports every month. Many possible causes have been hypothesized. Here are several possible correlates of theft which are reported on a monthly scale. See figure 10.1.

We might guess that car-theft is an outdoor activity, and may decrease with inclimate weather; alternatively, using a rational action based theory of crime, we might guess that when the desire for cars is strongest (namely cold or rainy weather), the crime rate should go up. It would be interesting, at least, to look at the effect of climate variables on auto theft rates. To access this, the high and low temperatures (averaged over a month) and the precipitation (total over a month) are recorded from NOAA data at Sand Point (Seattle).

Some theories of crime have postulated an economic rationale. As an indicator of the economic condition of the city, the Standard and Poor's[1]/Case-Schiller Home Price Index can be used. It has had a more or less monotonic increase over the time covered, which does reflect growing housing prices. What expectation to have of the effect of this variable is left to the reader's imagination.

Finally, a top-down theory of crime might indicate the leadership of the city would have some effect on the auto theft rate (either through enforcement regimes or indirectly through social programs). To include the possibility of this effect, the mayor of Seattle (who conveniently takes office January 1, leading to no ambiguous transition months) is coded. Over the span of these data there have been two Seattle mayors: Paul Schell and Greg Nickels.

The raw data are shown in table 10.1. The data span about 6.5 years and have no missing values.[2]

[1] What a great name for a finance company!

[2] Muggins here, who had to assemble the data in a spreadsheet, got bored in 2000 and stopped. The data actually extend further back for anyone who is inclined to retrieve it.

Month	Year	Auto theft	House Price	Precip. (in)	High Temp. (°F)	Low Temp.	Mayor
1	2000	858	100	3.65	46.6	35.4	Schell
2	2000	850	100.48	4.57	51.5	36.9	Schell
3	2000	772	102.24	2.86	52.4	37.7	Schell
4	2000	654	103.46	1.52	60	42.9	Schell
5	2000	563	104.78	3.52	61.1	47.1	Schell
6	2000	621	105.28	0.89	70.4	52.1	Schell
7	2000	604	105.82	0.22	74.6	55.6	Schell
8	2000	648	105.88	0.38	74.6	54.5	Schell
9	2000	617	106.1	1.59	70.4	52.2	Schell
10	2000	656	106.08	3.6	60.8	46.1	Schell
11	2000	806	106.33	3.53	50.2	36.8	Schell
12	2000	737	106.73	2.43	46.6	35.7	Schell
1	2001	764	106.7	3.05	48.3	36.8	Schell
2	2001	555	106.72	2.47	49	33.7	Schell
3	2001	672	107.41	2.82	52.5	38.7	Schell
4	2001	762	108.97	2.55	56.9	41.2	Schell
5	2001	735	109.76	1.34	64.9	45.7	Schell
6	2001	707	110.56	2.69	67.2	50.3	Schell
7	2001	728	110.84	0.74	73.7	54	Schell
8	2001	686	111.23	1.98	75.5	56.5	Schell
9	2001	694	111.72	0.43	69.5	54.4	Schell
10	2001	739	111.94	4.25	58	45.2	Schell
11	2001	877	111.97	9.4	52.9	42.7	Schell
12	2001	836	111.58	5.1	47.7	36.8	Schell
1	2002	696	111.79	5.68	45.7	37.6	Nickels
2	2002	579	112.07	4.43	49.4	35.9	Nickels
3	2002	551	112.74	2.68	48.8	36.9	Nickels
4	2002	576	113.4	2.79	56.6	41.1	Nickels
5	2002	552	114.18	1.34	61.6	45.7	Nickels
6	2002	575	114.84	1.36	71.5	52	Nickels
7	2002	717	115.28	0.7	74.5	55.7	Nickels
8	2002	747	115.49	0.18	76.7	55.4	Nickels
9	2002	730	115.62	0.65	71.6	50.8	Nickels
10	2002	853	115.88	0.51	59.3	44.9	Nickels
11	2002	851	115.87	2.86	55.1	41.3	Nickels
12	2002	881	116.18	5.24	48.7	39	Nickels
1	2003	871	115.8	6.74	51.5	40.1	Nickels
2	2003	808	116.27	1.68	48.6	36.4	Nickels
3	2003	711	117.04	5.11	53.9	41.1	Nickels
4	2003	763	118.23	2.72	56.5	42.9	Nickels
5	2003	796	119.23	1.32	64.1	46.8	Nickels
6	2003	855	120.15	0.95	72.6	52.9	Nickels
7	2003	680	120.84	0	79.9	56.4	Nickels
8	2003	638	121.88	0.3	78	56.2	Nickels
9	2003	677	122.42	1.62	72.7	53.5	Nickels
10	2003	676	123.43	6.98	62.3	48.7	Nickels
11	2003	803	123.53	5.65	49.6	37.7	Nickels
12	2003	774	124.38	4	47.2	37.4	Nickels
1	2004	728	124.42	7.14	45.3	36.7	Nickels
2	2004	744	125.22	2.45	52	37.7	Nickels
3	2004	656	126.33	1.8	55.8	40.5	Nickels
4	2004	604	128.23	0.64	63.8	43	Nickels
5	2004	716	130.15	2.23	66.2	49.4	Nickels
6	2004	652	131.94	0.62	73.5	53.2	Nickels
7	2004	686	133.27	0.4	79	58.2	Nickels
8	2004	851	134.21	3.05	79.9	59.8	Nickels
9	2004	804	135.22	1.94	67.4	52.8	Nickels
10	2004	1089	136.08	2.67	60.7	47.9	Nickels
11	2004	895	137.38	3.26	51.5	40.9	Nickels
12	2004	828	138.61	5.01	47	39.5	Nickels
1	2005	977	140.18	3.28	48.1	37.5	Nickels
2	2005	762	141.31	1.37	52	34.1	Nickels
3	2005	827	143.7	3.63	57.3	41.3	Nickels
4	2005	867	146.21	3.19	59.6	43.8	Nickels
5	2005	844	148.97	2.87	67.4	50.8	Nickels
6	2005	853	151.79	2.41	68.4	52.9	Nickels
7	2005	830	154.11	0.99	76.5	55.8	Nickels
8	2005	816	156.61	0.33	78.1	56.6	Nickels
9	2005	704	158.99	1.67	69	50.8	Nickels
10	2005	629	161.06	2.66	61.1	48.8	Nickels
11	2005	768	162.73	4.74	48.7	39.3	Nickels
12	2005	686	164.2	7.39	46.6	37.1	Nickels
1	2006	670	165.49	10.12	48.7	40.5	Nickels
2	2006	731	167.09	3.07	48.9	33.7	Nickels
3	2006	685	169.46	1.63	53.5	38.4	Nickels
4	2006	627	172.28	2.1	58.7	41.4	Nickels
5	2006	566	174.84	2.65	65.5	47.1	Nickels
6	2006	479	177.81	1.81	72.1	53.2	Nickels
7	2006	644	179.96	0.08	77.6	57	Nickels
8	2006	638	181.84	0.19	76.2	54.5	Nickels
9	2006	689	183.08	1.81	72.8	51.7	Nickels
10	2006	787	183.79	2.03	60.1	45.1	Nickels
11	2006	750	183.88	11.56	49.8	40.4	Nickels
12	2006	889	183.97	8	47.1	36	Nickels
1	2007	764	183.92	3.29	45.1	33.2	Nickels
2	2007	534	184.85	2.14	50.1	38.7	Nickels
3	2007	633	186.44	3.28	54.2	40.4	Nickels
4	2007	514	188.89	1.54	58.8	42.8	Nickels
5	2007	514	190.68	1.41	65.2	46.2	Nickels

Table 10.1: Raw data of auto theft for the past 6.5 years. Theft data are from the Seattle Police website. Housing index is a Standard and Poor's Case-Schiller index for Seattle, and climate data are from the NOAA station at Sand Point (in Seattle).

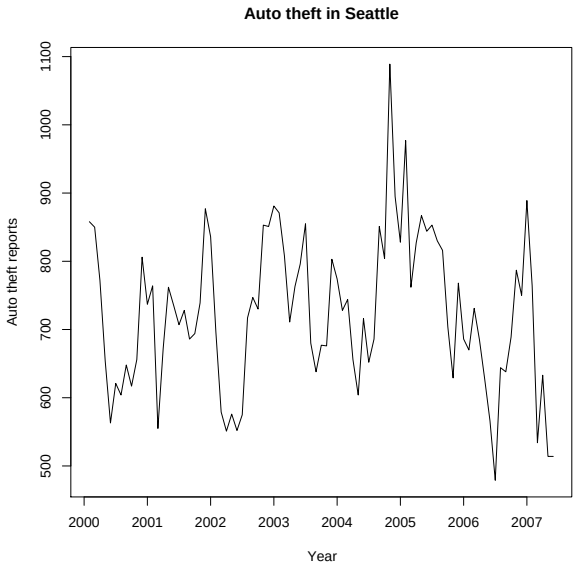


Figure 10.1: Auto theft reports from the Seattle Police for the past 6 years.

This data can be read into R and plotted (the plotting commands use some trickiness since time is recorded in different columns). The resulting plot is shown in figure 10.1.

```
> auto<-read.table(
      file='http://students.washington.edu/nesse/qerm514/data/autotheft.txt',
      header=T)
> plot(1:length(auto$precip),auto$atheft,type='l',
    main="Auto theft in Seattle",
    axes=F,
    xlab="Year",
    ylab="Auto theft reports")
> box()
> axis(2)
> axis(1,at=c(0,12,24,36,48,60,72,84),
    labels=c(2000,2001,2002,2003,2004,2005,2006,2007))
```

10.1.1 Thinking about the model

Model assumptions

Before blindly plunging in with the `lm()` command, it is worth reflecting a moment about whether a linear model is reasonable. Two immediate questions should be in mind:

- The data are counts, which may not necessarily be normally distributed.

- The data are time series, which may not be independent of each other.

The first of these–the normalcy of the errors–is a purely statistical question. Although a count must be an integer, for large counts, this does not matter as much. Furthermore, if the data were Poisson distributed for instance (a reasonable guess for count data of this sort) the Poisson approximates the normal for large values of the Poisson rate parameter λ. Since large λ would be needed to get large counts, it is a reasonable approximation that the data are normally distributed around their mean.[3]

The question of the time series aspect of the data is less a question of statistics. Recall that one of the assumptions of the ordinary linear model is the independence of observations. Data which occur sequentially through time are often (perhaps more often than not for real data) correlated. Various methods exist for assessing this correlation, however it really comes down to the process being modeled. In many cases such as populations, the response at one time should probably have an effect on the response at the next time (leading to so-called autocorrelation: correlation of a sequence with itself at a lag). Here, however, it is unreasonable to think that the auto theft rate in one month should influence the theft rate the next month, except as influenced by other variables (which themselves may be autocorrelated). Any sequentially ordered variables entered into a linear model should bring up the question of autocorrelation and whether a time series analysis is better justified.[4]

Interactions

There is one categorical variable in the model: mayor. Recall that categorical variables can have interactions with continuous variables. No interaction is interpreted as the categorical variable adding an amount to the response (which is dependent on the variable). Interactions can change the slope of the responses.

An interaction here would be interpreted as the mayor having an effect on the rain response (for instance, car thieves strike in the rain as a proportion to total under mayor Schell than Nickels). This seems to be unlikely. Thus with no further analysis, the interactions are ignored. (A better, although not convincing case could perhaps be made for an interaction between mayor and housing prices.)

Data issues

There are two things which present questions in this dataset. The first is the strong linear trend in the housing prices. The second is that there are two predictors for temperature which are likely to be very similar. See figure 10.2, plotted with the commands shown below.

```
> plot(1:length(auto$precip),auto$house,type='l',axes=F,
    xlab="Year",
    ylab="House price index",
    main="Housing prices through time")
> box()
> axis(2)
```

[3] This is a fairly intricate argument–one which is rarely thought through in most applications. Don't worry too much about it if it does not make good sense immediately.

[4] As it is, in fact, there *is* some level of autocorrelation, although not very much, in the data and residuals. We'll (unjustifiably) ignore it, as is commonly (but incorrectly) done, since it is not a topic for this class.

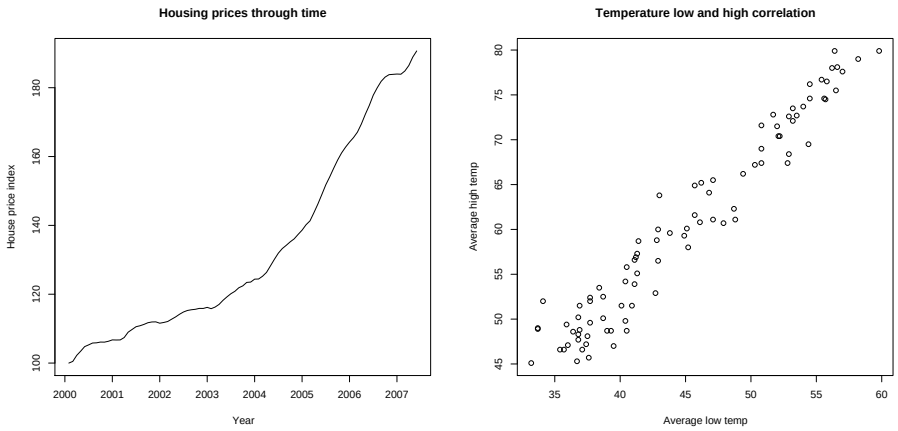


Figure 10.2: (Left) Housing index showing a strong linear trend, (Right) correlation between low and high temperature in the auto theft data.

```
> axis(1,at=c(0,12,24,36,48,60,72,84),labels=c(2000,2001,2002,2003,2004,2005,2006,2007))
>
> plot(auto$lo_t,auto$hi_t,main="Temperature low and high correlation",
    xlab="Average low temp",
    ylab="Average high temp")
```

In the first problem, the strong linear trend in the housing index, it may seem reasonable to detrend the data (that is, subtract out the line of best fit). This is a common procedure, but would be justified if fluctuations in the predictor would likely be more important the predictor itself. This is partially true: Some (though by no means all) of the increasing housing prices is due to inflation, rather than an actual increase in value. It could perhaps be justified on other grounds, but for this analysis will remain with a trend.[5]

With regard to temperature, it is often a good idea to combine similar variables into a more meaningful predictor. This reduces the number of parameters needed to fit while retaining the interpretability.[6] In this model, we have both a high temperature and a low temperature. It may seem unreasonable to include both in the model, since they are likely to predict the same things. Thus just using one of these, say high temperature, should probably be used instead of both. (An alternative approach is to use some combination of them, such as the average.)

10.1.2 The model

Using the reasoning above, a reasonable place to start on this model is with auto theft as the response, and precipitation, housing, high temperature, and mayor as predictors. The order here is

[5] A poor, though not unheard of mistake here is to argue that since the housing index has a trend but the response does not, the housing data should be detrended. This is generally poor reasoning in my view.

[6] There is an entire chapter in Faraway's *Linear models in* R on Shrinkage Methods, which are outside the scope of this course. We'll consider only the basic method of obvious combinations.

chosen more or less arbitrarily.

```
> auto.lm<-lm(atheft ~ precip+hi_t+as.factor(mayor)+house,data=auto)
> summary(auto.lm)

Call:
lm(formula = atheft ~ precip + hi_t + as.factor(mayor) + house,
    data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-215.156  -71.325    3.742   64.292  356.888

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            892.3028   120.0227   7.434 8.12e-11 ***
precip                  11.6570     6.6766   1.746   0.0845 .
hi_t                    -0.7479     1.4238  -0.525   0.6008
as.factor(mayor)Schell -50.2805    31.6555  -1.588   0.1160
house                   -1.0723     0.5176  -2.072   0.0414 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 107.9 on 84 degrees of freedom
Multiple R-Squared: 0.1218,     Adjusted R-squared: 0.08003
F-statistic: 2.914 on 4 and 84 DF,  p-value: 0.02609

> anova(auto.lm)
Analysis of Variance Table

Response: atheft
                 Df Sum Sq Mean Sq F value  Pr(>F)
precip            1  78321   78321  6.7211 0.01124 *
hi_t              1   4787    4787  0.4108 0.52331
as.factor(mayor)  1   2700    2700  0.2317 0.63154
house             1  50010   50010  4.2916 0.04137 *
Residuals        84 978850   11653
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

There are a few significant results in the parameters. Notice that precipitation is a moderate to poor predictor in using the t test, but is fairly significant in the F test. Recall that these differ by whether order matters: The F test is a sequential test, so for it the precipitation was compared against a mean of the response with no predictors. It is not surprising therefore that the precipitation (it it would have any effect at all) would show up as significant there. It is certainly an interesting relationship.

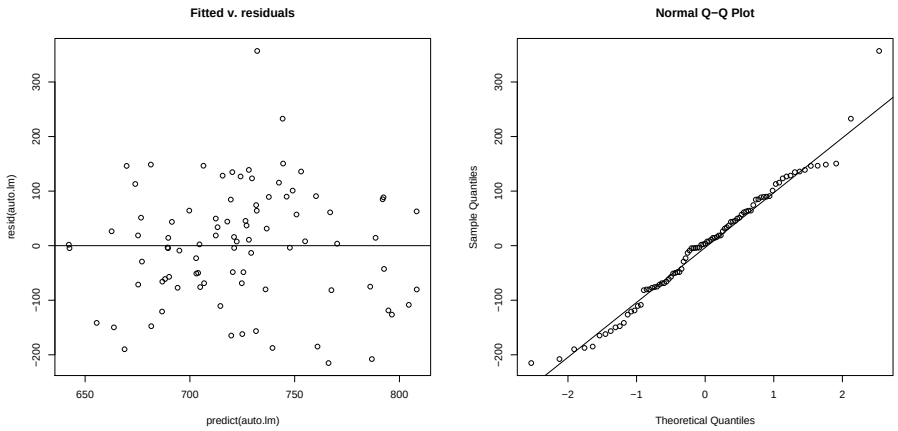


Figure 10.3: Fit of the auto theft model; the residual versus fitted model (right) shows no heteroskedasticity, while the qq plot shows an approximately normal distribution, consistent with model assumptions.

The housing index is the other predictor. Interpretation, however, should not immediately be considered causal. Since there was a strong trend in the housing data, it is reasonable to guess any trend in the auto theft would show up as a significant regression with the housing data.

Note that the precipitation has a positive response, meaning when it is raining a lot in a month, more cars are stolen. (Of course, as with housing, inference on the *causal* structure of the model should be done with caution.)

10.1.3 Diagnostics

Homoscedasticity

To ensure the errors are, in fact normally distributed with equal variance, as is assumed in the model, a few plots can be used here. A qq plot and a fitted v. residuals plot show the model assumptions hold, at least approximately. See figure 10.3.

```
> plot(predict(auto.lm),resid(auto.lm), main="Fitted v. residuals")
> abline(h=0)
>
> qqnorm(resid(auto.lm))
> qqline(resid(auto.lm))
```

Extreme observations

There is no real reason to expect extreme observations just from looking at the data, although with 89 observations they might be difficult to spot. Using first leverage then Cook's distance, it is possible to identify potentially problematic observations.

```
> lever<-influence(auto.lm)$hat
> length(lever[lever>10/89])
[1] 3
> sort(lever,decreasing=T)[1:5]
       83        23        73        84        46
0.2415073 0.1694823 0.1510996 0.1025875 0.1019686
>
> cd<-cooks.distance(auto.lm)
> length(cd[cd>.5])
[1] 0
```

The leverage shows three observations with high leverage (recall the rule of thumb that a leverage is high when it exceeds $2\frac{p}{n}$). The three points are 83, 23, and 73. The Cook's distance for these three points, however, is low so these points are not strong cause for concern. (Note that all the Cook's distances are below 0.5.)

This seems to be within the realm of reasonable variation; nothing that would immediately prompt corrective action (removing points, etc.). It is also possible to do a Box-Cox plot, however there is no indication that one would likely be needed (since the residuals are roughly normal to begin with).

10.1.4 Discussion

The two predictors of auto theft seem to stand out in the analysis: housing prices and precipitation. The coefficient was positive for precipitation, indicating more rain was associated with more theft, and negative for housing prices, indicating higher housing prices were associated with lower theft rates.

A common theme in discussion sections is to then come up with stories which are consistent with the data. For instance, more rain might actually cause higher theft since there are fewer around in the rain, thieves who are stealing for their own transportation will desire transportation more strongly in the rain, etc. It is more the topic of a philosophy class to determine whether these narratives actually constitute science, but they certainly extend beyond what the simple correlations indicate.[7] Keeping with the axiom that all of science is preliminary, however, it does indicate the sort of things which should be researched in the future.

10.2 Sexual dimorphism

Many animals vary morphologically by sex. Common traits which vary are coloration pattern, but size is one of the most dramatic traits to vary.[8] Mammals can vary pretty dramatically in size, although not nearly as dramatically as other classes of animals, see figure 10.4).

Several factors may influence sexual dimorphism. Two factors for which there are data are

[7] Necessarily, in fact. Some people would call those narratives "hypotheses" and hypotheses always contain more than the data can substantiate (the so-called underdetermination problem in the philosophy of science).

[8] The variation in size can cause all sorts of problems for identifying fossilized animals as well. Three species of moa, large flightless birds which went extinct on New Zealand, were recently shown to actually be one, highly dimorphic species though mtDNA extracted from ancient specimens. See Bunce et al, Nature 425:172-175.

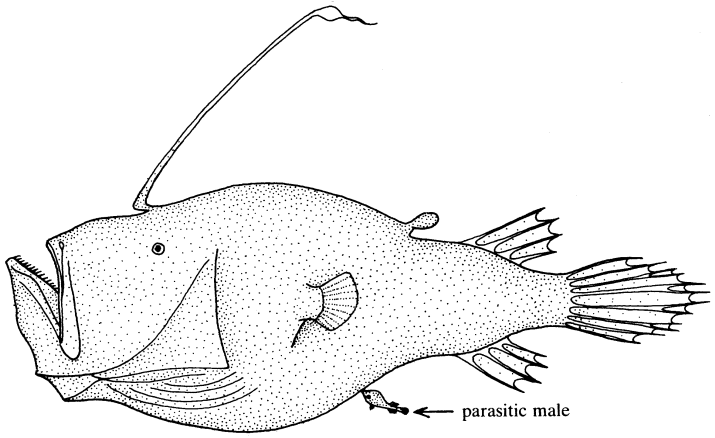


Figure 10.4: A drawing of the female (large) and male (small) individuals of the triplewart seadevil (*Cryptopsaras couesii*). According to Wikipedia, adult males become parasitic on females. The males eventually share blood supply with the females, and can be thought of as a sperm-generating organ of the female. An extreme example of sexual dimorphism. Picture by Dr Tony Ayling, licensed Creative Commons Attribution ShareAlike 1.0.

evolutionary history and mating system. The data come from Weckerly (1998)[9] and are available from the website under the name `dimorph.txt`.[10] Some of the data are incomplete, as often happens in practice.

First, import the data and remove missing observations.[11]

```
> dimorph<-read.table(
   file="http://students.washington.edu/nesse/qerm514/data/dimorph.txt",
   header=T)
>
> dimo<-NULL
> dimo$spp<-dimorph$spp[is.na(dimorph$mate)==F]
> dimo$male<-dimorph$male[is.na(dimorph$mate)==F]
> dimo$female<-dimorph$female[is.na(dimorph$mate)==F]
> dimo$mate<-dimorph$mate[is.na(dimorph$mate)==F]
> dimo$family<-dimorph$family[is.na(dimorph$mate)==F]
> dimo<-as.data.frame(dimo)
> names(dimo)<-names(dimorph)
```

The resulting collection is 222 species. Note that these data are not a random selection of mammals (it would be a fairly extensive list), but precisely those families which show a great deal of sexual dimorphism. This is important since it limits the sort of things which can be investigated with

[9] Weckerly, FW 1998. "Sexual-size Dimorphism: Influence of Mass and Mating Systems in the Most Dimorphic Mammals" *Journal of Mammalogy* 79: 33-52.

[10] I had to use optical character recognition to import the data from the paper, so there may be issues with accuracy.

[11] There may be more clever ways to deal with missing data, however this is sufficient for our purposes.

these data. Extrapolating from these data to a larger collection of families (such as all mammals) is going to be problematic.

10.2.1 Building a model

The male:female weight ratio is a reasonable starting place. Note however that the choice of M:F is arbitrary and could also be F:M. Although arbitrary, the decision may in fact be important since one or the other may be non-normal.

Possible predictors include mating system (see table 10.2) and taxonomic family. It might be possible to include something like male mass as a continuous predictor as well, since we might expect behavior of small mammals to be different from large mammals. Using components of the response as predictors can be tricky, however, since some combinations can be correlated simply because of the mathematics employed. (That's not to say it is not occasionally done.)

A simple model might use M:F as the response and family and mating behavior as the predictors. There are insufficient data to fit a model with interactions (many families have only one or two mating behaviors). Thus the model is limited to the additive model.

```
> dimo.lm<-lm(male/female~mate+family,data=dimo)
> par(mfrow=c(2,2))
> plot(dimo.lm)
> par(mfrow=c(1,1))
```

There are clear problems with this model. The fitted-residuals plot shows a cone shape, indicating heteroscedasticity. The qq-plot likewise strongly indicates non-normality. One option is to consider a box-cox transformation. Note that the arbitrary choice to choose M:F instead of F:M can show up in the box-cox plot.

```
> boxcox(dimo.lm)
```

The Box-Cox plot indicates a probable need for transformation. Possibly simply inverting the response, making this a females:males response would be sufficient. The fitted-residual plot looks a bit suspicious (see figure 10.7), although the qq plot is very much improved (although it still borders on suspicious).

```
> dimo.lm<-lm(female/male~mate+family,data=dimo)
> plot(fitted.values(dimo.lm),residuals(dimo.lm),
    main='fitted-residual plot',
    xlab='Fitted values',
    ylab='Residuals')
> qqnorm(residuals(dimo.lm))
> qqline(residuals(dimo.lm))
```

Not changing the response to more closely match the suggested Box-Cox transformation has preserved interpretability of the model, however the normality assumptions might be questionable. Nevertheless, we'll live on the edge and push forward.[12]

[12] Who said statistics were not exciting?

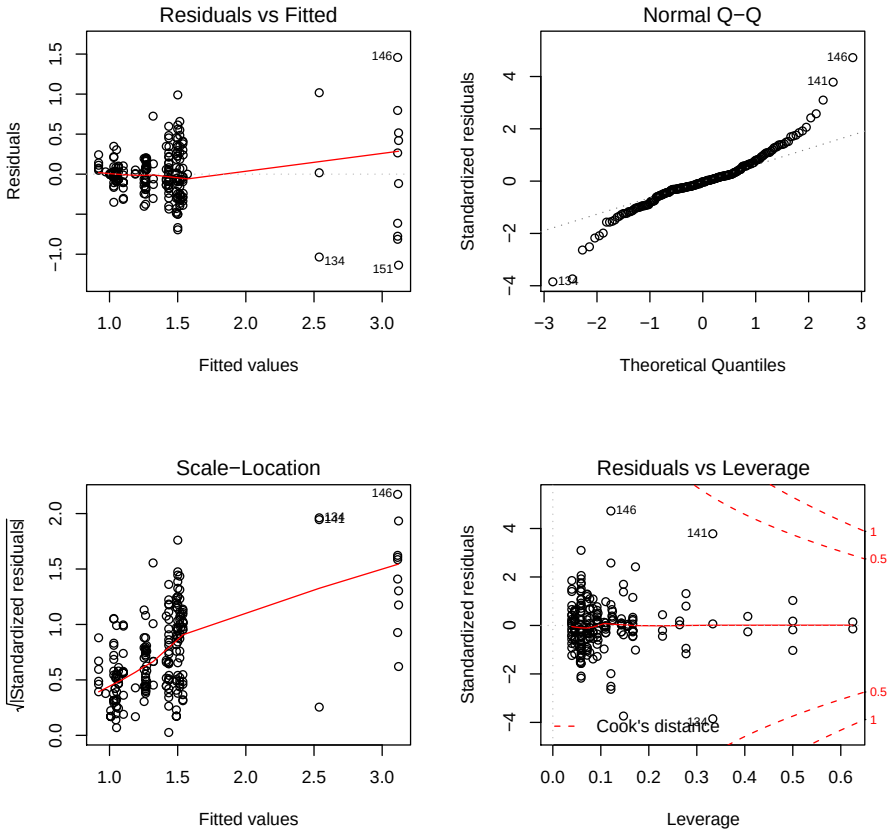


Figure 10.5: Diagnostic plots for the model using male:female weight ratio. Plots show definite problems.

A	Aquatic
D	Dominance rank competition
F	Tending
H	Harem
K	Multimale
M	Monogamy/polyandry
P	Polygynous
S	Mating season competition
T	Territorial

Table 10.2: Mating behaviors recorded for the sexual dimorphism data. For these purposes, the behaviors are taken to be mutually exclusive.

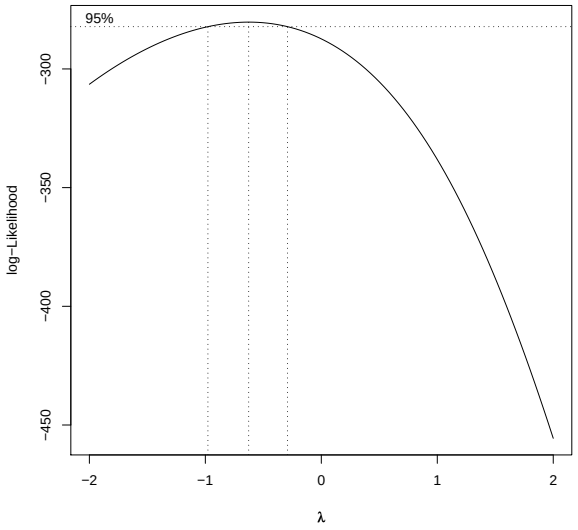


Figure 10.6: A Box-Cox plot of the male:female model.

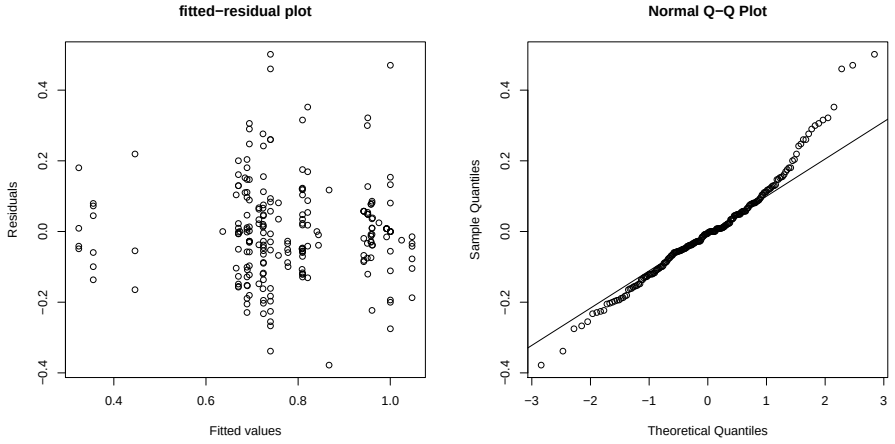


Figure 10.7: A fitted-residual plot using the F:M response (left) and qq plot (right).

10.2.2 Unusual observations

A check of the leverage reveals a few points with high leverage. Particularly, four points have a leverage of exactly 1. These points are highly influential as the sole representatives of their families. It is true that these points have a higher-than-normal influence compared with other points, however it would be unjustified to throw these points out.

Cook's distance can not be calculated for those four points (since Cook's distance involves dividing by $\sqrt{1-h_{ii}}$). The remaining points, however, have Cook's distances which are reasonable, indicating there are no other unusual observations.

10.2.3 Inference

Examining the model fits using F tests in an anova table indicates that both mating system and family are significant predictors.

```
> anova(dimo.lm)
Analysis of Variance Table

Response: female/male
           Df Sum Sq Mean Sq F value    Pr(>F)
mate        8 3.5476  0.4435 21.1280 < 2.2e-16 ***
family     19 1.9509  0.1027  4.8921 2.624e-09 ***
Residuals 194 4.0718  0.0210
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

It is also interesting to do t tests on the parameters themselves.

```
> summary(dimo.lm)

Call:
lm(formula = female/male ~ mate + family, data = dimo)

Residuals:
      Min        1Q    Median        3Q       Max
-0.377988 -0.076921 -0.004856  0.065367  0.501418

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.39502    0.17205   8.108 5.69e-14 ***
mateD                -0.71374    0.12721  -5.611 6.88e-08 ***
mateF                -0.60775    0.10749  -5.654 5.54e-08 ***
mateH                -0.55428    0.09279  -5.973 1.09e-08 ***
mateK                -0.55935    0.10448  -5.354 2.42e-07 ***
mateM                -0.38124    0.10819  -3.524 0.000531 ***
mateP                -0.33207    0.14900  -2.229 0.026984 *
mateS                -0.58434    0.12927  -4.520 1.07e-05 ***
mateT                -0.52286    0.10659  -4.905 1.97e-06 ***
```

```
familyBovidae         -0.06262    0.15354  -0.408 0.683833
familyCallitrichidae  -0.05336    0.16209  -0.329 0.742353
familyCebidae          0.03352    0.15831   0.212 0.832515
familyCercopithecidae -0.14657    0.14899  -0.984 0.326451
familyCervidae        -0.06355    0.15361  -0.414 0.679564
familyDaubentoniidae  -0.06295    0.23573  -0.267 0.789718
familyElephantidae    -0.20659    0.18503  -1.117 0.265572
familyGiraffidae      -0.15091    0.21194  -0.712 0.477312
familyHylobatidae     -0.05554    0.16457  -0.337 0.736109
familyIndriidae       -0.03836    0.19288  -0.199 0.842557
familyLemuridae       -0.07092    0.18058  -0.393 0.694938
familyLorisidae       -0.12041    0.19513  -0.617 0.537902
familyMacropodidae    -0.04731    0.15864  -0.298 0.765850
familyMustelidae      -0.20179    0.15887  -1.270 0.205543
familyOdobenidae      -0.72131    0.22492  -3.207 0.001569 **
familyOtariidae       -0.51683    0.15517  -3.331 0.001037 **
familyPhocidae        -0.39481    0.16729  -2.360 0.019264 *
familyPongidae        -0.15093    0.16374  -0.922 0.357810
familyTarsiidae       -0.10141    0.21230  -0.478 0.633422
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1449 on 194 degrees of freedom
Multiple R-Squared: 0.5745,     Adjusted R-squared: 0.5153
F-statistic: 9.703 on 27 and 194 DF,  p-value: < 2.2e-16
```

This seems to indicate that each mating system is significantly different from the aquatic mating system, and families (except for the three pinnipeds) do not differ significantly from the pronghorned antelope (the only extant member of family Antilocapridae). It would be a mistake, however, to indicate that this means mating systems are better predictors of dimorphism than family– remember that by choosing different base codings, different comparisons are significant. Comparison of models is really more the role of the F test. In fact, although interesting, it may be unnecessary to do a t test in this case since there is little value placed on the results (being that there are no baseline levels).

10.3 Discussion

There is good reason to think that family and mating system are important in determining the sexual dimorphism of species. Mating system is often identified as the causal factor which gives rise to evolutionary selective pressures for dimorphism, either through interspecific competition for mates (in polygamous mating systems), or displays in other systems. Thus this result is not surprising. Overall size may be fairly plastic, from an evolutionary standpoint, however there may be some evolutionary hold-overs which give rise to dimorphism where it is not expected by the mating system. Thus family may be a causal factor in that way.

Lecture 11

General Maximum Likelihood Estimates

11.1 Main ideas

- MLE method (review)
- Nonlinear models, a few examples
- Numerical optimization
- Asymptotic results
  - Likelihood ratio tests
  - Confidence intervals

This lecture is intended to give a reading knowledge of the topic; the details and subtleties here have been glossed over. Nevertheless these topics are widely used (and sometimes misused), so some understanding of the topic is important. Furthermore, the remainder of the material in this course relies on the results here, but the details have been worked out to assure the methods will work (usually). It is the only full lecture included here which was not in Loveday's original course.

11.2 Maximum likelihood

Recall that a model is a link between an observation response y_i and one or more predictors $x_1, \ldots, x_k$, together with some random components Λ (potentially a vector as well). The general expression for this is $y_i = g(x_1, x_2, \ldots, x_k, \Lambda)$. The often this is represented as Y_i being a random variable with a distribution f which depends both on the predictor(s) x_i and one or more of the distribution θ.

Thus observations can be thought of as draws from a random distribution. Since each observation is modeled to be independent, the joint distribution of all the observations is just

$$Y|x,\theta \sim \prod_{i=1}^{n} f(y_i, x_i, \theta) \tag{11.1}$$

Once the data are observed, this expression becomes a function of θ, known as the likelihood function. Since in many cases θ is unknown, if asked to come up with an estimate of θ the "best" estimate of θ, the natural choice is the one which maximizes this expression. That is[1]

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f(y_i, x_i, \theta) \tag{11.2}$$

In a few cases, this expression can be reasonably computed. However for numerical reasons, it is often convenient to take the log of equation 11.2. Note that since log is a strictly monotonic function[2], $\hat{\theta}$ can also be found using equation 11.3.

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log\left(f(y_i, x_i, \theta)\right) \tag{11.3}$$

For likelihood equations of one or two parameters, the log-likelihood equation can be plotted and the maximum visually described. See section 11.3.1 for examples of visualization. Many problems have multiple parameters (sometimes dozens or hundreds); maximum likelihood methods work there as well, however it is more difficult to visualize.

The goals of MLE analysis are the same as for any other modeling effort (see section 1.1). The estimates of the free parameters will have to be derived, and some measure of the quality of their fit is needed. Inference on the estimates such as hypothesis tests is useful, and assessment of model adequacy is valuable. Methods for doing these in general are roughly ranked in difficulty in this order: finding an estimate is the easiest, estimating its uncertainty is more difficult, inference is a bit more difficult than that, and general assessments of model adequacy are not covered here.[3]

Finding MLE estimates of parameters is done in one of two ways: analytically (which should be familiar, but is described briefly below), and numerically. Confidence intervals, at least under some conditions, can be described using an asymptotic distribution of the MLE. Inference in general is based on an asymptotic distribution of the likelihood ratio.

11.2.1 Finding the MLE analytically

It is often possible to maximize the log-likelihood expression analytically (that is, derive an explicit expression for it rather than rely on numerical methods). Such an approach was used in section 3.3.1 to derive the maximum likelihood estimates of β for the linear model.

The procedure relies on the following facts:

[1] Recall $\arg\max_{\theta}$ is read as "the argument θ which maximizes..." Here it simply means the θ that makes the expression which follows the largest.

[2] A strictly monotonic function always goes up as its argument increases.

[3] In fact, frankly, I don't know of any method of assessing model adequacy in general–only ones which apply to specific cases.

- At local maxima or minima on the interior of the parameter space, all first partial derivatives with respect to each parameter must be zero (provided they exist).
- The Hessian (the matrix of second derivatives) must be positive definite at minima and negative definite at maxima.

The first fact here is used to narrow down what points might be a maximum or minimum (Note the reverse of the first fact is not true: a point with zero first derivatives need not be either a max or minimum.) The second fact is sometimes called the second-derivative test, and is used to classify points as maxima or minima when needed.

Definition 11.2.1 (Critical point)**.** *A critical point of the function $y = f(x)$ is any point x_0 where $\frac{\partial y}{\partial x}|_{x_0} = 0$ (note this may be a vector equation if x is a vector).*

Relying on the first fact above, any critical point is a candidate to be the MLE. To find the critical points, one approach is to differentiate the likelihood, or more commonly (and equivalently) the log-likelihood with respect to each parameter and set each equation to zero. This will yield as many equations as there are parameters; solving these equations (often a non-trivial task) yields the candidates for the maximum likelihood estimate of the parameter(s).

If there is only one critical point (which is often the case), the second derivative test can be used to verify that the critical point is indeed the maximum. If there are multiple roots, the value of the log-likelihood expression at each point can be compared, indicating which is a global maximum.

Boundary values

The method of finding critical points will find only estimates on the *interior* of the parameter space. Formally, showing a point is a maximum requires showing that the maximum does not occur on the boundary of the parameter space as well.

In many cases, the boundary value is setting one of the parameters equal to a constant. In this case it is possible to simply substitute the constant for the parameter and find critical points using the remaining free parameters. Occasionally the boundary will be something more complex, generally written as $c = g(\theta_1, \theta_2, \ldots, \theta_k)$, where c is some constant and θ_1 through θ_k are the parameters. For this kind of problem, the method of Lagrange multipliers can be used to find maxima and minima.

11.2.2 Finding MLEs numerically

In many (if not most) problems, it will be difficult to find the critical points of a model. Thus the analytic method to find the MLE is difficult. However there is a large literature on numerical optimization (entire courses taught on it, in fact), and finding the MLE is really just finding the maximum of a function.

For most problems, there are too many parameters to numerically calculate a grid of possible parameter value combinations, as was done in section 11.3.1.[4] The likelihood surface (see figure 11.1 for example) too computationally complex to calculate every point of. Instead, some numerical method can be used.

One simple numerical optimization routine is simply to make a guess, and move in a direction which is upwards from that guess. These are known as quasi-Newtonian methods, which includes

[4]Sometimes this is possible, of course, and for as many as 3 or 4 parameter problems should probably be considered.

the BFGS (Broyden-Fletcher-Goldfarb-Shanno) method. It requires a starting value, and a function which is twice differentiable in all of the parameters.

The default option in R's `optim()` function is Nelder-Mead. It does not require differentiability, but is slower than BFGS. In most cases, the likelihood function should be twice differentiable and thus be amenable to BFGS.

There are several other options, including Conjugate-Gradient and Simulated Annealing. Covering the details of the optimization methods is outside of the scope of the course.

11.2.3 Important provisos on using the MLE

The maximum likelihood estimate (whether found analytically or estimated numerically) for parameters is a very powerful technique however several things should be kept in mind.

- There is no guarantee (in general) that the MLE will exist, or if it does, that it will be unique.
- As was seen for the estimate of the variance of an ordinary linear regression, the MLE may be biased or have other undesirable properties.
- In rare cases, the MLE will not converge (with increasing sample size) to the parameter at all.[5]
- The MLE may be difficult to calculate in practice.

11.2.4 Confidence intervals

Just as was done in section 4.3.1–developing confidence intervals around linear model parameters–it is also (sometimes) possible to develop confidence intervals around maximum likelihood parameters. The ordinary least squares parameters were fairly nice since the exact distribution of the estimated parameter was known. For maximum likelihood estimates, the estimator may not be analytic (it is solved numerically) or the transformation of the data is not easily worked into a probability distribution. In such cases, the asymptotic normality of the MLE is used.

Theorem 11.2.1 (MLE Normality)**.** *Suppose $\hat{\theta}$ is the root of the likelihood equation such that $\hat{\theta} \to \theta$ (this is a non-trivial but largely technical assumption that the MLE does, for large enough sample size, estimate the parameter[6]) and a few other regularity conditions, the estimate of a single parameter θ_i (possibly from a group of parameters θ) is approximately normal for large sample sizes: $\hat{\theta}_i^{(n)} \dot{\sim} N(\theta_i, I(\theta_i)^{-1})$, where $I(\theta_i)$ is shown below.*

$$I(\theta) = -E\left(\left(\frac{\partial^2 \mathcal{L}}{\partial \theta_i^2}\right)\right) \approx -\frac{\partial^2 \mathcal{L}}{\partial \theta_i^2}\Big|_{\theta=\hat{\theta}}, \tag{11.4}$$

where $\mathcal{L}$ is the joint log-likelihood of all the data.

[5] This may be worrying, but the one way of thinking about this: at its worst the MLE just delivers an estimate with no useful properties–the same as you'd get with a method of moments estimator or similar technique. The specifics of when an MLE exist are laid out in Casella and Berger.

[6] Formally, if the mle $\hat{\theta}^{(n)}$ is based on n samples, this requires $\lim_{n\to\infty} P(|\hat{\theta}^{(n)} - \theta| = 0) = 1$.

The proof of this result is rather technical and is omitted (but can be found in Lehmann and Casella's 1998 book *Theory of Point Estimation, 2nd ed.*). Notably, since theorem 11.2.1 is an asymptotic result, it is valid only for large sample sizes.[7] However it does give a method for approximating a confidence interval around a parameter estimate.

Using the following approximation, which only holds true for large samples, a confidence interval can be estimated.

$$\frac{\hat{\theta}_i - \theta_i}{\sqrt{I(\theta_i)}} \dot{\sim} N(0,1) \tag{11.5}$$

The rest of this development can be compared to section 4.3.1. The null hypothesis of a particular value of θ_i would be rejected if

$$\frac{|\hat{\theta}_i - \theta_i|}{\sqrt{I(\theta_i)}} \geq z_{1-\alpha/2}, \tag{11.6}$$

where $z_{1-\alpha/2}$ is the value such that $1 - \Phi(z_{\alpha/2}) = \alpha/2$ (recall $\Phi(x)$ is the cdf of the standard normal). Therefore the confidence interval for large samples is approximately

$$\hat{\theta}_i - z_{1-\alpha/2}\sqrt{I(\theta_i)} \leq \theta_i \leq \hat{\theta}_i + z_{1-\alpha/2}\sqrt{I(\theta_i)} \tag{11.7}$$

11.2.5 Inference

There are several methods of doing hypothesis tests using asymptotic distributions. The asymptotic distribution of the MLE, used above in equation 11.5 to test hypotheses, a procedure called *Wald's test*. A more common approach, however, is a likelihood ratio test based on the asymptotic distribution of the log likelihood ratio.[8]

Let $\lambda(X)$ be the likelihood ratio,[9]

$$\lambda(Y) = \frac{f(Y|\theta = \hat{\theta}_0)}{f(Y|\theta = \hat{\theta}} \tag{11.8}$$

Here f is the joint pdf of the data which depends on the parameters θ. The $\hat{\theta}_0$ is the MLE of θ under the null hypothesis, while $\hat{\theta}$ is the MLE under the alternative hypothesis.[10]

Intuitively, it should seem reasonable that the likelihood ratio is a good basis for a test. If λ is very small, the null hypothesis has a very low likelihood while the alternative is higher (λ is, after all, just the ratio of the likelihood functions). The reverse is also true. Words like "very small" or "large" are not a good basis for a test however. Quantifying these terms falls to Wilk's theorem.

[7] I don't know of a method other than simulation to determine, in general, if a given sample size is large enough in a given model.

[8] Just to add to the confusion, most commonly the term "likelihood ratio test" is applied to uses of this asymptotic distribution. However, strictly speaking, all the tests developed thus far are likelihood ratio tests (other, perhaps than the non-parametric tests), in that the decision rule is equivalent to rejecting the null when $\lambda < c$ for some constant c. Proving that a t or F test is, in fact, a likelihood ratio test is a fairly non-trivial problem, however.

[9] Note that whether the null is on top or bottom depends on the book used; This is consistent with Casella and Berger, but opposite of Michael Perlman's notes, who taught Stat 513 this year.

[10] This is maybe a bit confusing, since previously an MLE was an MLE, regardless of the hypothesis. In common formulations of a hypothesis test, the null hypothesis is a specific subset of a larger parameter space (such as, in a t test, $\mu = \mu_0$ under the null, where μ under the alternative can be anything not μ_0). Thus the MLE under the null is the maximum under the constraint that the parameters are in the null subset, while the MLE under the alternative requires the parameter be in the alternative space.

Theorem 11.2.2 (Wilk). *For Y a collection of n independent observations from the pdf $f(y|\theta)$, under sufficient regularity conditions $-2\log\lambda(Y) \to \chi^2_a$, where a is the difference between the number of free parameters in the alternative and the number of free parameters in the null hypothesis.*

Proof. (Sketch for a special case)[11] Suppose the null hypothesis under consideration is $\theta = \theta_0$, and the alternative is $\theta \neq \theta_0$. Since the parameter space under the null hypothesis is just a single point, θ_0, that is the maximum likelihood under the null. Under the alternative, it is the ordinary MLE, $\hat{\theta}$. Start by expanding the log likelihood around the mle (in θ) to the second order Taylor series:

$$l(Y|\theta) \approx l(Y|\hat{\theta}) + l'(Y|\hat{\theta})(\theta - \hat{\theta}) + \frac{l''(Y|\hat{\theta})}{2}(\theta - \hat{\theta})^2 \tag{11.9}$$

Note that since $\hat{\theta}$ is the MLE, $l'(Y|\hat{\theta}) = 0$. Evaluating at θ_0 and multiplying through by -2 yields

$$-2l(Y|\theta_0) \approx -2l(Y|\hat{\theta}) - l''(Y|\hat{\theta})(\theta_0 - \hat{\theta})^2. \tag{11.10}$$

Rewrite this as

$$-2\log\lambda(Y) \approx \frac{(\theta_0 - \hat{\theta})^2}{-l''(Y|\hat{\theta})^{-1}} \tag{11.11}$$

Note that by theorem 11.2.1, the denominator is asymptotically the variance of the MLE, this is approximately the square of a standard normal. Thus it is asymptotically distributed χ^2_1. □

11.2.6 Diagnostics

Since both confidence intervals and the likelihood ratio test rely on asymptotic distributions, there is a danger that the asymptotic approximation is not good enough for some particular application. One approach to judging whether a sample is large enough to use asymptotic approximation is to use Monte-Carlo trials. Repeatedly generate data under the null hypothesis (of the same size as the original data) and determine if the confidence intervals is reasonable, or if the likelihood ratio test works most of the time.

11.3 Examples

	Section	Description
Visualization	11.3.1	Commands to plot likelihood surfaces in one and two dimensions.
Zipf's law	11.3.2	A simple, one dimensional mle maximization.
Blue Crab	12.5.2	Three ways to calculate CPUE, including mle.

11.3.1 Visualizing the likelihood surface

Although rarely practical for applied problems (due to the large number of parameters in most "real" problems), graphical representations can give insight into lower dimensional problems which are useful in learning the methods.

[11] Honestly, I have never seen more than a sketch of this theorem.

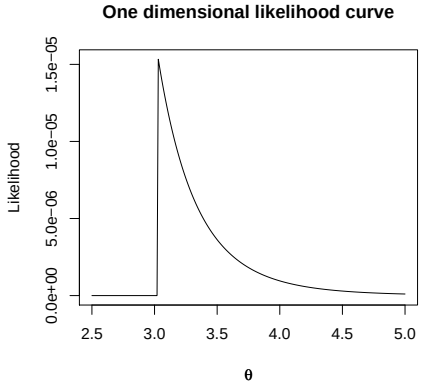
One dimensional likelihood

A one dimensional likelihood curve is usually the easiest to estimate. In fact, often there is little need to do the optimization since the maximum is usually pretty clear. Consider the case below, where $X \overset{\text{i.i.d.}}{\sim} \text{unif}(0, \theta)$. There are ten data, shown below.

1	2	3	4	5	6	7	8	9	10
0.85	1.19	1.83	2.91	0.65	2.87	3.02	2.11	2.01	0.20

Recall that a uniform probability density has the same probability of everywhere on the interval specified (here 0 to the unknown θ). Note that the likelihood that θ is smaller than the largest observation x (here point 7, with a value of 3.02) is zero. The likelihood is positive when $\theta >$ max.

```
> #Generate the data
> set.seed(1)
> x<-runif(10,0,3.2)
>
> #set up sequence of theta values to plot
> theta<-seq(2.5,5,.01)
> likeli<-NULL
> for(tt in 1:length(theta)){
      if(max(x)>theta[tt]){
          likeli[tt]<-0
      }
      if(max(x)<=theta[tt]){
          likeli[tt]<-1/theta[tt]^10
      }
  }
> plot(theta,likeli,type='l',xlab=expression(theta),ylab="Likelihood",
      main="One dimensional likelihood curve")
```



The mle can be reasonably guessed from the graph (and in fact, solved analytically), and is simply the largest observation of X. This is, notably, not unbiased (it will always be smaller than the actual value). It is also heavily (entirely) dependent on one observation, which does not make it an overly robust estimator. In that sense, the mle is not always the best choice for an estimator.

Two dimensional likelihood

Consider the case of 30 observations made from a normal(4,100) distribution (really high variance!).[12]

```
x<-rnorm(30,4,10)
```

The joint log likelihood of these observations is

$$\mathcal{L} = \sum_{i=1}^{30} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right) \tag{11.12}$$

$$= \sum_{i=1}^{30}\left(-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(x_i-\mu)^2\right) \tag{11.13}$$

This can be visualized in several ways. A wireframe is a clear way to show three dimensions, although it is hard to read actual values. A contour plot (isopleth plot) is an easier-to-read representation, but is less conceptual.

```
> ## Set up a matrix to plot, using a for loop
> likeden<-matrix(NA,length(sigma),length(mu))
> for(tt in 1:length(mu)){
      for(rr in 1:length(sigma)){
          likeden[rr,tt]<- sum(-1/2*log(2*pi)
                               - log(sigma[rr])
                               - 1/2*(x - mu[tt])^2/sigma[rr]^2)
      }
  }
>
> ## Plot the wireframe
> persp(sigma,mu,likeden,
      ltheta = -135, lphi = 75,
      phi=35,shade=1.25,
      xlab="sigma",ylab="mu",zlab="Log likelihood density",
      ticktype = "detailed")
>
> ## Plot the contour plot
> contour(sigma,mu,likeden,
      main="Contour plot of the log likelihood density")
```

11.3.2 Zipf's law

Biodiversity has long been a topic of fascination for ecologists. The word "biodiversity" is used in many contexts and has many definitions. One frequent definition is a quantification of the relative abundance within between taxa (how many individuals per taxon, usually species). Several

[12]Note that this is normal(4,100), yet the call in `R` is `rnorm(30,4,10)`. This is not a typo; recall that `R` uses mean and standard deviation, but the notation here is mean and variance.

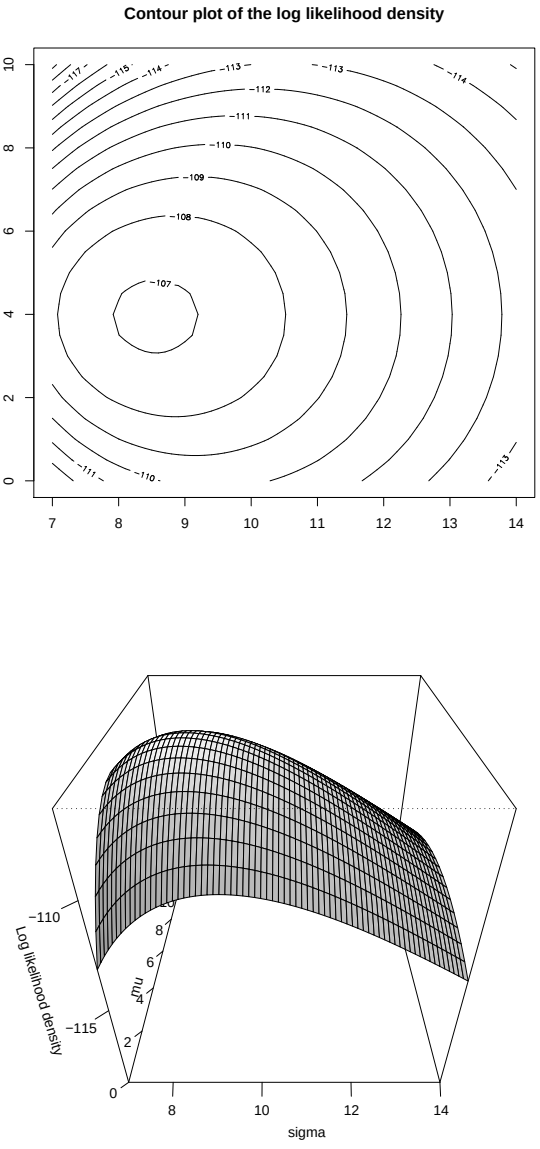


Figure 11.1: Contour (top) and wireframe (bottom) plots of log likelihood equation for 30 normal(4,100) data points. Despite the high variance, the mle estimate is still pretty good.

measures have been developed to assess this form of biodiversity. Two of the common methods (which are not dealt with here, but are worth mentioning) are Shannon Biodiversity index and Simpson's index.

Definition 11.3.1 (Biodiversity Indices)**.** *Both the Shannon Biodiversity index and Simpson's index examines the within taxon counts of individuals c_i, to the total among all groups N. Define the proportion of the individuals within taxon i to be $p_i = c_i/N$. The Shannon index, often written H' is defined below.*

$$H' = -\sum_i p_i \log(p_i) \tag{11.14}$$

Simpson's index D is similarly defined.

$$D = \sum_i p_i^2 \tag{11.15}$$

Neither of these are going to be used here, however. Instead, Zipf's distribution will be fit. Zipf's distribution has been occasionally fit to species abundance curves–it predicts an exponential fall off in diversity with higher ranks. Specifically, if r_i is the rank of the taxon (it is the r_ith most abundant taxon), Zipf's law predicts the abundance of the taxon should be

$$z_i = \frac{A}{r_i^k} \tag{11.16}$$

where $A = \left(\sum_i \frac{1}{i^k}\right)^{-1}$ is a normalizing constant (to ensure the probabilities sum to 1). Thus Zipf's law gives expectations of the probabilities for each taxon. The probability mass function for data could be modeled as multinomial, with the Zipf's probabilities as the arguments. The goal of this exercise is to fit Zipf's constant k to a dataset. Note that this fit constant is interpretable as another index of biodiversity (like Shannon's or Simpson's).

The dataset for this example comes from petfinder.com, particularly the dog listings. There are about 300 observations of dog breed (along with other information we won't use here). Breeds are mostly listed according to their American Kennel Club name (with a few exceptions) and for mixed individuals, the first breed noted was listed.[13]

The results are shown in the table below.

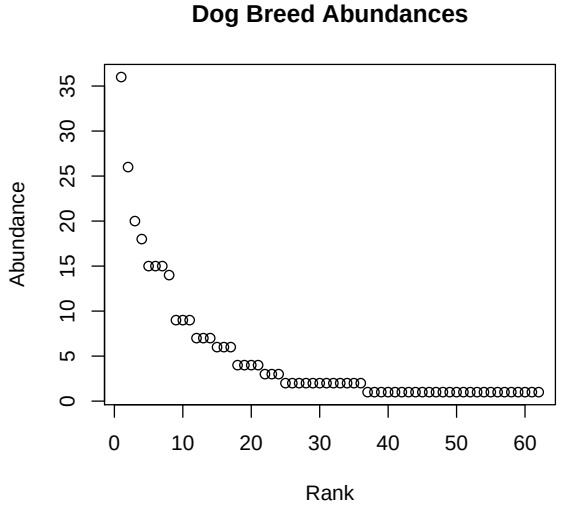
Airedale Terrier	7	American Bulldog	1	American Eskimo Dog	2	American Staffordshire Terrier	2
Australian Cattle Dog	4	Australian Shepherd	20	Basset Hound	4	Bichon Frise	1
Bloodhound	1	Border Collie	14	Boston Terrier	1	Boxer	1
Brittany	1	Bull Terrier	1	Cairn Terrier	1	Carolina Dog	1
Chihuahua	18	Chinese Crested	2	Chow Chow	1	Cocker Spaniel	1
Collie	3	Dachshund	7	Dalmatian	3	Doberman Pinscher	2
English Springer Spaniel	1	French Bulldog	1	German Shepard Dog	6	German Shepherd Dog	15
Golden Retriever	1	Great Dane	2	Great Pyrenees	2	Greyhound	1
Italian Greyhound	9	Japanese Chin	1	Labrador Retriever	26	Lhasa Apso	1
Manchester Terrier	1	Mastiff	2	Miniature Pinscher	4	Pekingese	2
Pembroke Welsh Corgi	1	Pit Bull Terrier	36	Pointer	1	Pomeranian	9
Poodle	2	Pug	2	Rat Terrier	1	Rottweiler	6
Schipperke	4	Shar Pei	15	Shih Tzu	1	Siberian Husky	15
Staffordshire Bull Terrier	2	Tibetan Mastiff	1	Tibetan Terrier	1	Unknown	7
Welsh Corgi	1	West Highland White Terrier	6	Wire Fox Terrier	3	Xoloitzcuintli	2
Yellow Labrador Retriever	9	Yorkshire Terrier	1				

All the observations can be found online (with additional data on the age (baby, young, adult, and senior), sex, and shelter). First, download the data and create an ordered list of abundances.

[13]There are several aspects about these data which make them difficult to analyze. Since they are all rescue dogs, it is rare the actual breed or breed composition in known. The shelter may list a popular name over a more accurate name for identifiability. Moreover, since some shelters specialize in a particular kind of dog, that breed may be overrepresented here.

Unlike other datasets, the delimiter between fields here are commas, so use the `read.csv()` command rather than `read.table()`.

```
> dogs<-read.csv(
    file="http://students.washington.edu/nesse/qerm514/data/dogs.csv",
    header=T)
> biod<-sort(as.numeric(table(dogs$breed)),decreasing=T)
> plot(biod, main="Dog Breed Abundances", xlab="Rank",ylab="Abundance")
```



We'll investigate two methods for fitting these data to the Zipf curve: maximum likelihood estimate of a multinomial, and a linear fit to log transformed data. (In general it will not be possible to do both, of course, and the two methods do make different assumptions, but it is useful here for comparison.)

Fitting using maximum likelihood methods, first set up the likelihood function for the Zipf parameter k and counts X. A reasonable approach is to model X_i as from a multinomial with probability z_i, given by Zipf's distribution. Let N be the total number of dogs surveyed, here 300, and B be the number of breeds represented, here 62.

$$f(X|k) = \binom{N}{x_1 x_2 \cdots x_B} z_1^{x_1} z_2^{x_2} \cdots z_B^{x_B} \tag{11.17}$$

In this case, it is probably easier to maximize the log likelihood. The optimization method will use the `optim()` function, which minimizes. Thus the best approach may be to minimize the negative log likelihood function. This is coded and run below.

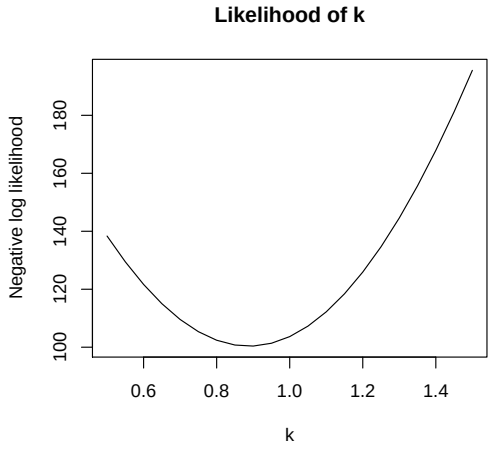
```
> ## making a zipf function
> zipf<-function(n,k){
    x<-1:n
    H<-sum(1/x^k)
```

```
    (1/x^k)/H
  }
>
> #Describing the negative log likelihood
> bioc<-function(k){
      probs<-zipf(length(biod),k)
      return(-log(dmultinom(biod,sum(biod),probs)))
  }
>
> #Optimizing
> optim(fn=bioc,par=list(k=1),method="BFGS")
$par
        k
0.8883557

$value
[1] 100.3615

$counts
function gradient
      17        3

$convergence
[1] 0

$message
NULL
```

The result of the optimization is shown in the `par` section: $k = 0.888$. The `value` given is the value of the negative log-likelihood being fit, while the `counts`, `convergence` and `message` are for displaying information about the fitting routine.

This optimization could also be done graphically. A plot of the likelihood is shown below.

```
> k<-seq(.5,1.5,.05)
> likeli<-NULL
> for(tt in 1:length(k)){
      likeli[tt]<-bioc(k[tt])
  }
>
> plot(k,likeli,main="Likelihood of k",xlab="k",
     ylab="Negative log likelihood", type='l')
```



Even though the possible models we could specify are endless, there is no guarantee Zipf's distribution is the right one (in fact, the data are not a particularly good fit to Zipf's distribution here). However absent other information, the ordinary diagnostics are more difficult to interpret.

Fitting the model using OLS

In the above, the data were modeled as multinomial. A more common approach[14] would be to fit a linear model to the log transformed data. In particular, if the observations can be thought of as approximately Nz_i, and $z_i = Ar_i^{-k}$, then observations $x_i \approx ar_i^b$, for some fit constants a and b. Modeling this formally, assume $x_i = ar_i^b\eta_i$ where η_i are independent log-normal errors.[15] Thus taking the log of both sides yields

$$\log(x_i) = \log(a) + b\log(r_i) + \epsilon_i \tag{11.18}$$

where ϵ_i are normally distributed. This is, of course, an ordinary linear regression.

```
> lr<-log(1:62)
> log.biod<-log(biod)
> bio.lm<-lm(log.biod~lr)
> summary(bio.lm)

Call:
lm(formula = log.biod ~ lr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92959 -0.15935  0.02299  0.17302  0.46435

Coefficients:
```

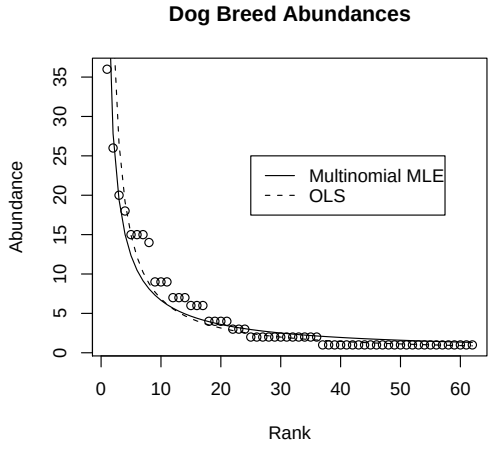
[14] I would be willing to bet nearly universally, in fact

[15] A log-normal distribution is one wherein the log of the log-normal random variable is normally distributed.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.51311    0.12571   35.90   <2e-16 ***
lr          -1.12441    0.03811  -29.51   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2687 on 60 degrees of freedom
Multiple R-Squared: 0.9355,     Adjusted R-squared: 0.9345
F-statistic: 870.7 on 1 and 60 DF,  p-value: < 2.2e-16
```

Note that $b = -k$, so the fit here is not too far off the Zipf's fit: 0.888 compared to 1.1244. Comparing the fit of the actual data, the curves are nearly indistinguishable.

```
> plot(biod, main="Dog Breed Abundances", xlab="Rank",ylab="Abundance")
> lines(1:length(biod),zipf(length(biod),.8883557)*sum(biod))
> lines(1:length(biod),I(exp(4.51311)*(1:62)^-1.12441),lty=2)
> legend(25,25,legend=c("Multinomial MLE","OLS"),lty=c(1,2))
```



Lecture 12

Nonlinear least squares

12.1 Main ideas

- Nonlinear models
- Fitting routines
- Confidence intervals

There are a lot of derivatives in this section. Note that the notation $\dot{g}$ means the derivative with respect to the argument of the function g.

12.2 Nonlinear models

Often the models being fit in ecology are not lines, or forms which can be transformed into lines. The function which links the predictors and the response can take on any number of forms (see examples). Nonlinear least squares is a reasonable way of generating some estimates for a wide range of models, and pretty good estimates for a more narrow range.

Suppose that the response y_i is a (known) function g of predictors X_i given parameters θ; that is $y_i = g(X_i, \theta)$. Note that this function, unlike those previously, does not include a random component (that will be considered later). Just as in least squares minimization in linear regression, the goal is to minimize the sum of squared residuals. Denote the nonlinear least squares estimate of θ with a tilde, $\tilde{\theta}$.

$$\tilde{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - g(X_i, \theta))^2 \tag{12.1}$$

This equation is homologous to equation 3.7, replacing the g with a linear function (and relabeling the parameters as β rather than θ. Just as in that section, the least squares procedure does not formally assume a particular error. However, as will be shown later, the nls estimate has some nice properties if the model (with a random component) is

$$Y_i = g(X_i, \theta) + \epsilon_i, \tag{12.2}$$

where ϵ_i are independent and each distributed $N(0, \sigma^2)$. Note that when the predictor does not have a subscript, that is X, the function g gives the expected value of all the observations y. Thus g is a vector valued function.

Under this model, the nonlinear least squares estimate can be interpreted as the MLE for the parameters (this argument follows the same pattern as for the linear case, in section 3.3.1). As the MLE, the confidence intervals from the previous lecture (asymptotic distribution of MLE) can be applied.

When is this (equation 12.2) reasonable? One way of thinking about the model is that the model is capturing observational errors, but not other sorts of errors. For example, perhaps the sole source of error in the model is the measurement of the response; this would make a good model for this purpose.

Of course, reality is generally more complex than this model is capable of capturing. In these cases, nls does produce some estimate of the parameters. However in the interpretation of those parameters it is important to keep in mind the simplicity of the nls model, when compared to reality. Adding confidence intervals, for instance, or developing hypothesis tests akin to Wald's test or the Likelihood ratio test may be categorically unjustified if the model itself is wrong.

12.3 Fitting methods

The in lecture on the general MLE method one section, section 11.2.2, discussed numerical optimization techniques. These would work here, however may be less efficient than those which take full advantage of what is known about the model being fit. In particular, starting from an estimate of θ_0 (note this is NOT the θ under the null hypothesis), the estimate can be improved using Newton's method.

Linearizing g gives a way to guess at better values of θ from a starting point θ_t.

$$g(x|\theta) \approx g(x|\theta_t) + \left.\frac{\partial g}{\partial \theta}\right|_{\theta=\theta_t} (\theta - \theta_t) \tag{12.3}$$

Note that $\left.\frac{\partial g}{\partial \theta}\right|_{\theta=\theta_t}$ is (at least potentially) a matrix, as is $\theta - \theta_t$. Using the normal, additive error assumed in the model, this is simply a linear model, with the unknown parameter $\theta - \theta_t$. Thus the least-squares solution is already known from section 3.3. That is, the value of $\theta - \theta_t$ which minimizes $(y - g(y|\theta))'(y - g(y|\theta))$ is approximately the value which minimizes

$$[(y - g(X|\theta)) - \dot{g}(X|\theta_t)(\theta - \theta_t)]' [(y - g(X|\theta)) - \dot{g}(X|\theta_t)(\theta - \theta_t)] \tag{12.4}$$

which is just a linear regression model (with $\dot{g}$ as the model matrix and $y - g(y|\theta)$ as the response). The best fit $\theta - \theta_t$ is therefore $(\dot{g}(y|\theta_t)'\dot{g}(y|\theta_t))^{-1}\dot{g}(y|\theta_t)'(y - g(y|\theta_t)$. This suggests that θ_{t+1} should be $\theta_t + (\dot{g}(y|\theta_t)'\dot{g}(y|\theta_t))^{-1}\dot{g}(y|\theta_t)'(y - g(y|\theta_t)$.

Using this method, the estimate of θ generally improves. It still requires a starting value, and in some cases, will not find the true maximum. As with many local search methods, the Gauss-Newton method can also get stuck on local maxima rather than global.

12.4 Confidence intervals

Confidence intervals around parameters can be developed using the asymptotic properties of the MLE (see section 11.2.4). The usual provisos apply: the distribution only really applies to very large

samples, since it is based on asymptotic properties. Furthermore, as was noted in the introduction to this lecture, the nls model is often used for expediency rather than accuracy. Thus the confidence intervals based on that model may be wildly inaccurate.

Never-the-less, it is sometimes necessary to develop confidence intervals around parameter estimates, even with the problems mentioned. To do this, it is necessary to find the asymptotic variance, which by theorem 11.2.1 is $-\frac{\partial^2 \mathcal{L}}{\partial \theta_i^2}\Big|_{\theta=\hat{\theta}}$, which simplifies to $\sigma^2 \dot{g}(X|\theta)'\dot{g}(X|\theta)$ (where σ^2 is the variance of the error ϵ_i). The variance for a particular parameter j in the parameter vector θ is just the jjth element of this matrix times σ^2.

The estimate of σ^2 is given by $(y - g(X|\hat{\theta}))'(y - g(X|\hat{\theta}))/(n-p)$ for n observations and p estimated parameters. Although this is really only valid for large samples, negating the need for a t distribution, some sources (eg Seeber and Wilde's *Nonlinear regression*) will still set up confidence intervals using a t_{n-p} distribution. That is, the key fact for use in the confidence interval is

$$\frac{\hat{\theta}_i - \theta_i}{\sqrt{\hat{\sigma}^2(\dot{g}(X|\theta)'\dot{g}(X|\theta))_{ii}}} \dot{\sim} t_{n-p} \tag{12.5}$$

The rest of the development of the confidence intervals follows the development in section 4.3.1.

12.5 Examples

	Section	Description
Dogfish	2.8.5	An example of nonlinear least squares in a randomization test.
Simple NLS	12.5.1	A simple sinusoidal regression to detect the orbit of the Earth.
Blue Crab	12.5.2	Three ways to calculate CPUE

12.5.1 Simple NLS

The study of cycles in ecology have been a topic of great interest for more than a century. People who studied animal populations noticed that there were regularities to their rises and falls. Numerous explanations have been devised for these, most commonly interactions with prey.

A noted ecologist Dr. Hairbrained T. Sciuridwatcher has collected some data on the abundance of ground squirrels in New Jersy, which he believes show some strong cycling behvaior. He has just published an extensive monograph detaining the cycles and speculating on their meaning and cause.

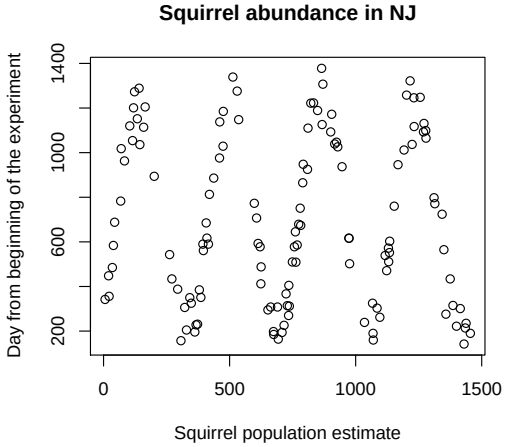
The data (made up for this example) are available on the website.[1] Data on populations should, unquestionably, be treated as a time series (certainly knowing the population on one day influences the population on the next.) Nevertheless, you can take a quick and dirty look at the data using nonlinear least squares.

First import the data and plot it.

```
> sq<-read.table(
      file="http://students.washington.edu/nesse/qerm514/data/squirrel.txt",
```

[1] `http://students.washington.edu/nesse/qerm514/data/squirrel.txt`

```
      header=T)
>
> plot(sq$day, sq$abund)
> plot(sq$day, sq$abund,
      main="Squirrel abundance in NJ",
      xlab="Squirrel population estimate",
      ylab="Day from beginning of the experiment")
```



These results do seem to indicate a cyclic pattern. Nonlinear least squares can give some basic information about the pattern. Specifically, it is possible to fit the model $y = A\sin(\frac{2\pi}{T}(d-\phi)) + B$ to the data. The fit parameters, A, T, ϕ and B have the interpretation of amplitude, period, phase shift, and mean respectively.

```
> sq.nls<-nls(abund ~ A * sin(2*pi/T * (day - phi))+ B,
+     start=list(A=400, T= 350, phi=0,B=800))
> summary(sq.nls)

Formula: abund ~ A * sin(2 * pi/T * (day - phi)) + B

Parameters:
    Estimate Std. Error t value Pr(>|t|)
A    512.864      9.819   52.23   <2e-16 ***
T    363.710      1.031  352.88   <2e-16 ***
phi   53.506      2.252   23.76   <2e-16 ***
B    717.332      7.360   97.46   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 78.13 on 126 degrees of freedom
```

```
Number of iterations to convergence: 7
Achieved convergence tolerance: 2.463e-06
```

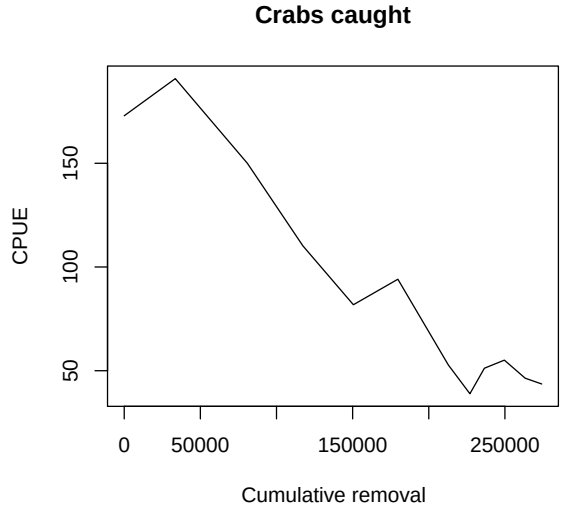
All the fits are significant, using the asymptotic t distribution. As far as interpretation, note that the period is very close to 365 (days). Hey, it happens to the best of us.

12.5.2 Blue Crab CPUE

This is a classic problem in fisheries:[2] As fishing removes animals, it becomes increasingly difficult to catch the remaining animals. As the stock declines, it takes more work and time to catch the few animals which remain, in general, at least on a per animal basis. This effect can be shown for data on the blue crab.

The standardization of catch by effort needed to catch fish is usually termed "catch per unit effort," or CPUE. Effort is measured in a variety of ways, such as pots set, soak time, area swept by a trawl net, hooks set on a longline fishery, etc. Below is the results of twelve weeks of fisheries data plotting the cumulative catch (total removal) prior to that week, against the CPUE for that week.

```
> crab<-read.table(
      file="http://students.washington.edu/nesse/qerm514/data/crab.txt",
      header=T)
> plot(crab$rem,crab$cpue,type='l',
      main="Crabs caught",
      xlab="Cumulative removal",
      ylab="CPUE")
```



[2] taken straight out of Loveday's notes.

These results can be looked at using three tools developed so far: linear regression, maximum likelihood estimation, and nonlinear least squares.

A common method for modeling the catch data is the Leslie model. In it, the probability of catching any one animal is related effort via the equation $p_t = 1 - e^{-cf_t}$, where c is a (fit) catchability coefficient and f is the known effort. Assuming independent captures, the number caught is a function of the number in the population N, after removal of x_t individuals up to time t. Thus the total number of individuals in the population in time t is $N - x_t$. (This model assumes no substantial growth in the population.)

The virgin biomass (N), as it is called, is unknown of course, but of great interest. Each week's observations do give some information about N. In particular, under the independence assumption, each week's catch n_t could be modeled as a binomial mass function.

$$f(n_t|N, p_t) = \binom{N - x_t}{n_t} p_t^{n_t} (1 - p_t)^{N - x_t - n_t} \tag{12.6}$$

For simplicity, p_t is left as it is, although in fact (as noted above) p_t is modeled as a function of effort, $p_t = 1 - e^{cf_t}$T. he joint observations of all the catches n over K weeks is therefore

$$\prod_{t=1}^{K} \binom{N - x_t}{n_t} p_t^{n_t} (1 - p_t)^{N - x_t - n_t}. \tag{12.7}$$

This model can be fit three ways:

- Directly, using maximum likelihood methods
- Approximately, using nonlinear least squares
- Even more approximately, using linear techniques.

Here all three methods will be applied, however the focus is on the nonlinear least squares technique.

Maximum likelihood

There are two parameters which need to be fit: N, the virgin biomass, and c, the catchability coefficient. Since N is so large, it will be computationally very difficult to calculate the binomial in equation 12.7. However a normal is a good approximation to the binomial for large N (really N does not have to be anywhere near this big–a few dozen is usually enough). Using the fact that the binomial is approximately equal to the normal for large N, it is possible to show

$$\binom{N - x_t}{n_t} p_t^{n_t} (1-p_t)^{N - x_t - n_t} \approx \frac{1}{\sqrt{2\pi(N - x_t)p_t(1 - p_t)}} \exp\left(-\frac{1}{2(N - x_t)p_t(1 - p_t)} (n_t - (N - x_t)p_t)^2 \right). \tag{12.8}$$

Thus the maximum likelihood estimator can be found by optimizing this function. Two issues are likely to come up in R in performing this optimization: First is the scaling of c and second is the choice of initial conditions. Since the expected value of c is very small, the default step size is an order of magnitude larger, so the optimization is likely to fail immediately. The fix this, the input value of c is divided by 10^6 in the function. The second problem is more difficult, and is often encountered in these sorts of problems. The range of initial parameter values which give log

likelihoods other than ∞ (due to rounding of the machine digit accuracy) are a fairly small range. Thus had a guess of the solution not been known apriori, finding those initial conditions would be rather difficult.

```
> crabPMF<-function(zz){  ##zz[1] = N, zz[2] = c
      f<-crab$eff
      n<-crab$samp
      x<-crab$rem
      N<-zz[1]
      c<-zz[2]/1000000
      cvar<-(N-x)*(1-exp(-c*f))*exp(-c*f)
      muc<-(N-x)*(1-exp(-c*f))
      res.tmp<- -1/2*log(2*pi) -1/2*log(cvar) - 1/(2*cvar)*(n-muc)^2
      return(-sum(res.tmp))
  }
>
> optim(fn=crabPMF,par=c(334000,4),method="BFGS")
$par
[1] 293310.8304    847.1344

$value
[1] 18125.47

$counts
function gradient
     107      100

$convergence
[1] 1

$message
NULL

Warning messages:
1: In log(cvar) : NaNs produced
2: In log(cvar) : NaNs produced
```

Even with those adjustments, there was still a warning that the estimated variance got close to zero. These warnings may indicate the solution is unreliable. The final answer using maximum likelihood directly is, for N, 293310.83, while the final c is $847.1344 \times 10^{-6} = .000847$. However, as noted, the optimization had difficulty and these results may not be accurate.

Nonlinear least squares

Note that the optimization, after making the normal approximation, similar to a nonlinear least squares model. The main difference is that here, the variance is assumed to be constant, while in the maximum likelihood, the variance scales with $N - x_t$. Nevertheless, the model is much easier to fit and solve.

```
> nls(samp~(N-rem)*(1-exp(-c*eff)),start=list(N=300000,c=.005),data=crab)
Nonlinear regression model
  model:  samp ~ (N - rem) * (1 - exp(-c * eff))
   data:  crab
       N         c
3.286e+05 6.101e-04
 residual sum-of-squares: 157400870

Number of iterations to convergence: 5
Achieved convergence tolerance: 9.592e-08
```

The fitting method is also fairly robust to selection of different starting values. The final values it arrives at are $N = 328600$ and $c = .0006101$. This is a bit different than the mle, although both are similar to an order of magnitude.

12.5.3 Linear fit

The linear method relies on the approximation of $p_t = 1 - e^{-cf_t}$ by its first order Taylor expansion in c, $p_t \approx -cf_t$. Thus in particular, $n_t = (N - x_t)p_t \approx (N - x_t)(-cf_t)$. Dividing both sides by f_t yields

$$cpue_t = \frac{n_t}{f_t} = cN - cx_t \tag{12.9}$$

This has the appearance of a linear equation, and thus can be solved via the ordinary least squares model $\beta_0 + \beta_1 x_t$. Thus $cN = \beta_0$ while $-c = \beta_1$. Solving these yields $c = -\beta_1$ and $N = -\beta_0/\beta_1$. Running the ordinary least squares is fairly simple, however.

```
> crab.lm<-lm(cpue~ rem,data=crab)
> summary(crab.lm)

Call:
lm(formula = cpue ~ rem, data = crab)

Residuals:
    Min      1Q  Median      3Q     Max
-19.117 -12.647   3.685   9.875  24.241

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.854e+02  9.312e+00   19.91 2.25e-09 ***
rem         -5.613e-04  4.886e-05  -11.49 4.40e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.99 on 10 degrees of freedom
Multiple R-Squared: 0.9296,     Adjusted R-squared: 0.9225
F-statistic:   132 on 1 and 10 DF,  p-value: 4.396e-07
```

Solving for N gives $-\frac{185.4}{-.0005613} = 330304.6$ while $c = 0.0005613$. This approach is computationally the easiest, and for all intents and purposes, gives an answer which is very similar to the nonlinear least squares. Confidence intervals for these estimates could be derived using proagation of error techniques,[3] see Loveday's notes for more details.

[3]Also known as the delta method.

Lecture 13

Generalized linear models

13.1 Main ideas

- Count regression
- Link functions
- Bernoulli response
- General form of the glm
- Deviance

13.2 Count regression

Measured responses in ecology are not always continuous (or even approximately continuous) variables, much less normally distributed (as is assumed for ordinary regression). One of the most common non-normal responses is counts. Although in previous examples have relied had count data as a response (see the auto theft example in section 10.1 for instance), these were approximated by a normal distribution. This approximation is not always very good, and often requires *post hoc* transformations such as Box-Cox to apply a normal regression model.

A much cleaner approach, when dealing with count data, is to throw out the assumption of normal error and use something else. Two commonly encountered error functions are the Poisson and the binomial.

The Poisson distribution, defined in definition 0.3.17, is commonly used for describing counts over a fixed period of time. The probability of getting a response $y \in \{0, 1, 2, \dots\}$ is given by the pmf shown below, which depends on the rate parameter λ.

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!} \tag{13.1}$$

The expected value of the a Poisson random variable EY is λ. The parameter λ is often called the "rate parameter," and intuitively, larger values of λ mean generally larger values of Y.[1] Note that λ really only makes sense if it is positive. (If $\lambda < 0$ the probability of getting an odd y is negative, which would be disconcerting.)

Recall that a binomial(N, p) random variable has a support of $\{0, 1, \ldots, N\}$ and assigns probability to outcome y by the following formula.

$$\binom{N}{y} p^y (1-p)^{N-y} \tag{13.2}$$

The binomial distribution already came up in this course, when looking at the χ^2 test (section 2.5). There the binomial was thought of as a series of N independent trials with two possible outcomes, with p of the probability of success on any one trial. The binomial is the number of successes in the N trials.[2] In fact, the χ^2 test and the binomial glm are closely related.

13.2.1 Model structure

In ordinary linear regression, the normal error ϵ_i was added on to the end of the predictor. Unfortunately, this simple structure is not the form of most generalized linear models. The general form will be discussed in the next lecture, however for Poisson and binomial data, the predictors can be used to estimate the parameter (λ for the Poisson and p for the binomial), through a link function discussed below.

Poisson

To model the response Y as a Poisson random variable, it is necessary to specify a λ. Since the goal of the regression analysis is to develop a model which links the response Y to the predictors X, perhaps making λ a function X is a reasonable starting point.

Absent other considerations λ_i (the rate parameter for the ith observation) could be a linear combination of the predictors for that observation X_i, with fit parameters β, that is $\lambda_i = X_i\beta$. This can be fit, but it has a problematic issue: λ *must* be positive while $X_i\beta$ could, at least in theory, be negative. Thus it may make more sense to fit some invertible function g.

$$\lambda_i = g^{-1}(X_i\beta) \tag{13.3}$$

The function g is called the *link function*. (Conventionally, g maps E$Y = \lambda \to X\beta$, so the function of $X\beta$ is, under this convention, g^{-1}. This is purely a convention and, had the development of the topic been different, gone the other way.[3]) The link function g, for the Poisson model, should take positive numbers and map them to all real numbers, that way the g^{-1} will force λ to be positive.

[1] Some mental images which might help picture this: the number of people arriving at a movie theater in a ten minute interval, the number of raindrops on a roof in a 5 second interval in a rainstorm, or the number of lightning strikes in Washington State in a year. With some assumptions (of varying realism), these could be modeled with a Poisson random variable.

[2] Some examples of binomial random variables, Y might be the number of heads in N flips of a coin, where each flip has probability p.

[3] Or more likely, been left to the author, leading to every author developing their own notation which results in the confusion that is modern statistics today.

Name	$g(p)$	$g^{-1}(X\beta)$	Description
Logit	$\log\left(\frac{p}{1-p}\right) = X\beta$	$p = \frac{\exp(X\beta)}{1+\exp(X\beta)}$	Sometimes called the log-odds.
Probit	$\Phi^{-1}(p) = X\beta$	$p = \Phi(X\beta)$	Recall that $\Phi(x)$ is the standard normal (that is $N(0,1)$) cumulative distribution function. Since a CDF always returns a probability, it ensures the p is between 0 and 1.
Cauchit	$\tan\left(\pi(p-\frac{1}{2})\right) = X\beta$	$p = \frac{1}{\pi}\arctan(X\beta) + \frac{1}{2}$	Formally this can be described as the CDF of the Cauchy distribution, although it is perhaps unnecessary to do so.
C. Log-Log	$\log(-\log(1-p)) = X\beta$	$p = \exp(-e^{X\beta})$	Complimentary log-log.

Table 13.1: Possible choices for the link function for a binomial glm in R.

Two choices for a link function (which are set in R) are $g(\lambda) = \log\lambda$ and $g(\lambda) = \sqrt{\lambda}$. Note that these are equivalent to $g^{-1}(X\beta) = \exp(X\beta)$ and $g^{-1}(X\beta) = (X\beta)^2$. Plugging these into equation 13.3 forces λ_i to be positive. (A further discussion of choosing link functions is included below)

The Poisson glm model is therefore to model each response Y_i as a Poisson random variable, with rate parameter $\lambda_i = g^{-1}(X_i\beta)$. As with ordinary linear regression, the β are free parameters which are fit to all observations.

Binomial

The binomial has two parameters, p and N, however most commonly, N is known.[4] Modeling each response Y_i as binomial, it is necessary to estimate a p_i for that observation. As with the Poisson case, since the goal is to link the response Y_i with the predictors X_i, a reasonable starting point is

$$p_i = g^{-1}(X_i\beta) \tag{13.4}$$

for some invertible link function g. The link function is specified so that $p_i = g^{-1}(X_i\beta)$ is between 0 and 1 (as a probability must be). Some of the options for g are specified in table 13.1.

Thus the binomial glm models each observation Y_i as binomial(N_i, p_i) where N_i is known and $p_i = g^{-1}(X_i\beta)$ is a fit parameter depending on the predictors for that observation, X_i.

13.2.2 Choosing a link functions

Choosing a link function is one of the least satisfying parts of using a generalized linear model. Most often it will be difficult to determine the appropriate link function (other developments in linear modeling such as quasi-likelihood and generalized additive models, which are not part of this course, will make this a bit less arbitrary), resulting in more or less arbitrary or conventional decisions.[5] However there are some natural choices in certain cases.

[4] Fitting N for even the simplest case is difficult.

[5] So what do you do? For myself, I do two things: 1. generally stick to the default link function, and 2. Make sure my interpretation of the resulting fit would be the same for any link function.

Log link

In the Poisson glm, the default link function is $\lambda = \exp(X\beta)$. One way of thinking about this is to argue that λ is the product of multiple factors, each of which is $e^{x_{ij}\beta_j}$.

Logit

The odds of an event (probability of occurring over the probability of not occurring) is arguably just as natural an expression of uncertainty as probability. Thus the logit link has the same product interpretation for odds in the binomial glm as the log link has in the Poisson glm.

Probit

Probit models have some history in statistics, even predating generalized linear models by almost 40 years. Suppose the predictors, $X_i\beta$ give some measure of the tolerance;[6] if the tolerance is exceeded, then the outcome is a failure, if it is not exceeded, a success. Under these circumstances, it can be shown that the probability a randomly drawn trial (from a normal distribution) is below tolerance follows a probit model.

13.3 Inference

In the ordinary linear regression problem, the sum of squared residuals was used to assess model fit and perform hypothesis testing. In fact, two similar measures exist for generalized linear models. The first such measure, the generalized Pearson's X^2 is has a familiar look–sum of squared residuals over a variance estimate–while the other assessment statistic, deviance, may look familiar from the discussion of maximum likelihood.

13.3.1 Pearson's goodness of fit

Definition 13.3.1 (Pearson's X^2)**.** *The Pearson's X^2 statistics is a measure of model fit (or discrepancy) which takes the form*

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{Var(y_i)}. \tag{13.5}$$

(The notation X^2 may be confusing, since X is also used for the model matrix. This is done for consistency with other texts (Faraway, Dobson, McCullagh and Nelder for example), however is not ideal. Hopefully the context will make the notation clear.)

As will be shown below, certain cases of the Pearson statistic will reduce to tests which have already been encountered in this course.

Theorem 13.3.1 (Binomial GLM - χ^2 relation)**.** *In a binomial glm, the Pearson X^2 statistic is equivalent to the usual χ^2 test statistic*

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{13.6}$$

[6] This description largely follow's Faraway's discussion in *Extending the linear model with R*. The particular example he uses is student's aptitude: if the student's aptitude exceeds a question's difficult, the student gets the answer correct. The tolerance distribution is made up of aptitudes of all students.

Proof. This is a fairly involved proof (although not conceptually difficult) for large numbers of responses. For simplicity, it will be shown here only for the simple case of 1 observation. Note that O_i is y_i for successes and $n_i - y_i$ for failures, while $E_i = n_i\hat{p}_i$ for successes and $n_i(1-\hat{p}_i$ for failures. In the χ^2 contingency table, the calculation is over both successes and failures, thus

$$\frac{(y-np)^2}{np} + \frac{(n-y-n(1-p))^2}{n(1-p)} = \frac{(y-np)^2(1-p) - p(-y+np)^2}{np(1-p)} \tag{13.7}$$

$$= \frac{(y-np)^2}{np(1-p)} \tag{13.8}$$

Now note that the variance of a binomial observation is $np(1-p)$, thus proving the result. □

In fact, even when this is not the usual χ^2 test statistics, asymptotically Pearson's X^2 is still χ^2_{n-p} distributed, where p is the number of parameters estimated and n is the number of observations. This result is strongest when n is large for binomials or the λ is large for Poisson glms (both of which should result in fairly large counts).

13.3.2 Deviance

The second goodness of fit measure in generalized liner models is termed deviance. It also generalizes the idea of a sum of squared residuals, however in a radically different way than Pearson. Recall main results from the sum of squares was the comparison of models (usually simple model to the model beign fit) via ratios of sums of squares.

Rather than comparing to a very simple model, as was done in the ordinary linear regression problem, the model in generalized linear models is compared to a model with the same number of parameters as points being fit, termed a *saturated model*.[7] A saturated model has the same link function and error distribution as the model being fit.

The definition here will rely on the dispersion parameter ϕ, which will only become important in the next lecture, discussing the general form of glms. The dispersion parameter for both the Poisson glm and Binomial glm is taken to be 1.

Definition 13.3.2 (Deviance)**.** *The deviance*[8] *of a model is negative twice the log likelihood ratio of the model being fit to the saturated model, times the dispersion parameter. That is*

$$D = -2\phi \log\left(\frac{f(y|\theta)}{f(y|\theta_{sat})}\right) \tag{13.9}$$

The negative twice log likelihood ratio by itself is termed the scaled deviance.

When comparing nested models, deviance gives a good way to assess model fit, via the following theorem (compare to theorem 4.4.1 on F-tests for ordinary linear models). Again, this is stated in general using a dispersion parameter which will not become important until the next lecture.

Theorem 13.3.2 (Deviance tests)**.** *Comparing a larger glm with deviance D_{large} to a smaller, nested glm (with the same error and link function) with deviance D_{small} can proceed in two ways:*

[7] It is also sometimes called a maximal model or full model. Note, however, that these notes used the term "full model" to denote a model with all interaction terms, see section 6.3.1.

[8] These definitions follow Faraway's definitions.

If the dispersion is not estimated for the model being considered, the difference in deviances is χ^2 distributed. For models which do estimate a dispersion parameter, $\hat{\phi}$, the ratio of the difference of deviances divided by their degrees of freedom to the estimated dispersion parameter is approximately F distributed. That is

$$D_{small} - D_{large} \quad \dot{\sim} \quad \chi^2_{p_{large}-p_{small}} \tag{13.10}$$

$$\frac{(D_{small} - D_{large})/(p_{large} - p_{small})}{\hat{\phi}} \quad \dot{\sim} \quad F_{p_{large}-p_{small},n-p_{large}} \tag{13.11}$$

where $\hat{\phi} = X^2/(n-p)$, using Pearson's X^2 to estimate the dispersion parameter.

This theorem is has a very similar appearance to the F tests previously discussed. This is not coincidental; in fact, it is the homologue. The deviance of a Gaussian model (an ordinary linear model) is identical to the sum of squared residuals. Thus one way of thinking of deviance is the non-normal equivalent to the sum of squared residuals.

13.3.3 Distribution of fit parameters

The parameters are fit using maximum likelihood, so it should come as no surprise, (recalling the asymptotic normality of mles, theorem 11.2.1) that the parameter estimates are asymptotically normal. Thus using the information estimates, it is possible to derive asymptotic standard errors and confidence intervals, or devise t (or more appropriately z) tests for the null that specific parameters equal zero, just as was done for the ordinary least squares regression.

13.3.4 Interpreting `R` output

Just as with ordinary multiple regressions, `R` provides two standard methods for interpreting the output of a glm: the summary table (given, as before, with the `summary()` command) and the analysis of deviance table (anodev, given by the `anova()` command). Interpretation is very much like section 4.5, except that the distributions are asymptotic.

By default, the `anova()` command does not perform a test on glm fits. However, one can be specified using the `test=` parameter. See examples in this and the next lecture for more details.

13.4 Examples

Name	section	description
Simple Poisson regression	13.4.1	Poisson regression, for reference on the commands.
Simple Binomial	13.4.2	Binomial regression, for reference on the commands.
Bartlett's Cuttings	13.4.3	Binomial regression with interactions.
Insect spray	13.4.4	Poisson regression for an experiment with categorical predictors.

13.4.1 Simple Poisson

A response which has a Poisson error structure should be count data, so they should appear to be non-negative integers. The predictors may take the form of any variable seen thus far: continuous or categorical. Below is 10 observations of made up data, with a continuous predictor `x1` and categorical predictor `x2`, and response `res`. All of these data can be found in the `simppois.txt` datafile online.

```
> sim<-read.table(
+     file="http://students.washington.edu/nesse/qerm514/data/simppois.txt",
+     header=T)
> sim
   res        x1 x2
1    3  0.852092  a
2   14  6.802208  b
3    6  0.768770  a
4   18 10.228164  b
5   13  5.111556  a
6   10  5.791032  b
7   14  9.655684  a
8    9  8.327390  b
9    4  1.722764  a
10  10  3.053211  b
```

The regression is fairly straightforward to run, using similar syntax to the more familiar `lm()`. Now the family of the regression must also be specified, however. Terms like interactions are still able to be added using the same syntax as before.

```
> # No interactions
> sim.glm<-glm(res~x1+x2,data=sim,family=poisson)
>
> # Interactions
> sim.glm<-glm(res~x1*x2,data=sim,family=poisson)
```

This command uses the default link function, which for Poisson is the log. To change link functions, say to square root, the command is inserted into the `poisson` portion.

```
> sim.glm<-glm(res~x1+x2,data=sim,family=poisson(link=sqrt))
```

To check diagnostics or residuals, the same commands work as for `lm()`. Since there are several types of residuals of interest, this can be specified in the `residuals()` command. For instance, the Pearson residuals can be found using the command. Note that raw residuals are called `response` residuals in `R`.

```
> residuals(sim.glm,type="pearson")
```

The commands to get inference on the model are, as before, `summary()`, for tests on the parameters, and `anova()` for comparisons of the model. Most often, the `anova()` command will be the first one used. By default, `anova()` does not output a test, however a test can be specified (even the wrong test, more on this in the next lecture). To perform the χ^2 test, for instance, use the command shown below.

```
> anova(sim.glm,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: sqrt

Response: res

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     9    22.2337
x1    1  15.6147         8     6.6190    0.0001
x2    1   0.2291         7     6.3899    0.6322
```

The resulting table is commonly called the *analysis of deviance* table. The test is to compare the differences in deviance between models (in the `Deviance` column) to a χ^2 distribution on the difference in degrees of freedom (noted in the `Df` column). These results indicate `x1` is a significant predictor while `x2` is not (at least after `x1` has entered the model).

13.4.2 Simple Binomial

There are two common methods for fitting the binomial response. Binomial responses should have two categories: alive/dead, male/female, etc. Information from both of these needs to enter the model, unlike other glms which have only one variable as a response. For instance, the data below are made up binomial data with one predictor.

successes	failures	total	predictor
32	33	65	-0.7206097
35	10	45	0.9339525
72	17	89	1.1083952
49	24	73	0.3340350
88	10	98	1.0452074
45	24	69	-0.2159193
59	35	94	-0.0784096
56	45	101	-0.5929465
73	16	89	0.8988237
51	41	92	-0.3855277

First read in the data.

```
> sib<-read.table(
    file="http://students.washington.edu/nesse/qerm514/data/simpbinom.txt",
    header=T)
```

First method

In a sense, both the successes and failures are the response, so both can be grouped into a single $n \times 2$ matrix. That matrix, then, becomes the response.

```
> res<-cbind(sib$success,sib$fail)
```

The `cbind()` command creates a matrix out of two vectors by appending them as columns. The regression can now be run.

```
> sib<-glm(res~sib$x,family=binomial)
```

Now all the commands can be run as usual.

```
> summary(sim.glm)

Call:
glm(formula = res ~ x1 + x2, family = poisson(link = sqrt), data = sim)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.44818  -0.64584   0.07391   0.51410   1.23804

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.08902    0.29601   7.057  1.7e-12 ***
x1           0.17925    0.05355   3.347 0.000816 ***
x2b          0.16925    0.36014   0.470 0.638383
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22.2337  on 9  degrees of freedom
Residual deviance:  6.3899  on 7  degrees of freedom
AIC: 52.804

Number of Fisher Scoring iterations: 4
```

Second method

A more intuitive way of thinking about the regression is to use the proportion of the total which are successes enter as the response. This, however, loses information (1/5 should be less important, for instance, than 10/50). The totals are therefore entered as weights in the regression. To use this method, first set up a collection of proportions (success/total).

```
> probs<-sib$success/sib$total
```

Now the regression can be run using the `probs` as the response, and the `total` as weights.

```
> sib.glm<-glm(probs~sib$x,weights=sib$total,family=binomial)
```

Note that the model fit is identical to the model above.

```
> summary(sib.glm)
```

```
Call:
glm(formula = probs ~ sib$x, family = binomial, weights = sib$total)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9748  -0.6224  -0.2017   0.4301   1.8258

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.65546    0.07889   8.308  < 2e-16 ***
sib$x        0.91270    0.12064   7.566 3.86e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.846  on 9  degrees of freedom
Residual deviance:  6.441  on 8  degrees of freedom
AIC: 56.106

Number of Fisher Scoring iterations: 4
```

13.4.3 Bartlett's cuttings

These data[9] were used to illustrate an older method (although still commonly used) contingency table approach to analysis of counts. Here, the same data will be used to illustrate a generalized linear model. The data are the number of plants which are alive or dead after planting a cutting, as a function of two categorical variables: the time of planting (early or late)[10] and length of cutting (long or short). Since the response (dead/alive) is binary, the count of alive and dead plants is a binomial response.

Time	Length	alive	dead
early	long	156	84
early	short	107	133
late	long	84	156
late	short	31	209

There are two models which might be examined here (at least which involve both predictors): a model with no interactions and a model with interactions. Note that the model with interactions, however, is fitting a unique parameter to every cell. (Of course, the link function can also be varied to generate new models as well, however they should have minimal impact on the interpretation of the results.)

First, import data. A bit of rearranging is needed to put the data into a form which can be used in `glm()`.

[9] I took the data from Bartlett's 1935 paper on contingency tables. Bartlett himself got the data from another source. Forgive me for not looking up the original source here, *The journal of pomology and horticultural science*, 1934. It is in the library, but is not, unfortunately, online. It turns out pomology is the study of fruit. Who knew?

[10] Originally termed "at once" and "in spring"

```
> bart<-bsp<-read.table(
+    file="http://students.washington.edu/nesse/qerm514/data/bart.txt",
+    header=T)
> y<-cbind(bart$alive,bart$dead)
> x1<-bart$time
> x2<-bart$length
```

Now fit the additive model, and check the significance of the predictors. Since the error has only one parameter, the variance is known so the χ^2 test on deviances can be used.

```
> bart.glm<-glm(y~x1+x2,family=binomial)
> anova(bart.glm, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     3    151.019
x1    1   97.579         2     53.440 5.175e-23
x2    1   51.147         1      2.294 8.572e-13
```

Both of these models seem to fit very well. We could also examine the individual parameters with the `summary()` command.

Another model which might be fit to these data is to include interaction terms between the predictors. Thus, for instance, if long cuttings did well in the early planting, but short did well in the late planting, the interaction term would detect that. Note, however, that the data for fitting this model is at a limit: fitting the interaction term fits one parameter for every data point in the model. Such a model is termed "saturated."

```
> anova(bart.glm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                     3    151.019
x1     1   97.579         2     53.440  5.175e-23
```

```
x2     1   51.147          1      2.294  8.572e-13
x1:x2  1    2.294          0 -8.371e-14      0.130
```

Here, as expected, the residual deviance drops to zero (up to a round-off error). The addition of an interaction term proves to be not significant.

13.4.4 Insect sprays

In this example, which is one of the datasets built into R,[11] the number of insects killed[12] by trials of different insecticides in an agricultural experiment. The results are shown below.

count	spray	count	spray	count	spray	count	spray
10	A	3	D	17	B	1	E
7	A	5	D	17	B	1	E
20	A	12	D	19	B	3	E
14	A	6	D	21	B	2	E
14	A	4	D	7	B	6	E
12	A	3	D	13	B	4	E
10	A	5	D	0	C	11	F
23	A	5	D	1	C	9	F
17	A	5	D	7	C	15	F
20	A	5	D	2	C	22	F
14	A	2	D	3	C	15	F
13	A	4	D	1	C	16	F
11	B	3	E	2	C	13	F
17	B	5	E	1	C	10	F
21	B	3	E	3	C	26	F
11	B	5	E	0	C	26	F
16	B	3	E	1	C	24	F
14	B	6	E	4	C	13	F

One approach to analysis might be to transform these data so that the response is more normal. The Box-Cox transformation might be used (in fact, the square root works pretty well), although the motivation for it is not clear.[13] A more convincing approach might be a generalized linear model with a Poisson error.

The model being fit is to model each count as Poisson(λ_j) where $\lambda_j = \exp(X\beta)$, and X is the usual model matrix for categorical variables (using the default `contr.treatment` coding). The parameters of interest are, as before, β. There is no need to read in the data since it comes with R.

```
> spray<-glm(count~spray,data=InsectSprays,family=poisson)
> anova(spray,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log
```

[11] Although the reference to the original work is incorrect in the R help file. The correct citation is Beall, Geoffery. 1942. "The transformation of data from entomological field experiments so that the analysis of variance becomes applicable." *Biometrika* 32(3-4):243-262.

[12] Well, frankly, it is not clear if the data reported are the number of insects killed or the number left alive. We'll assume they mean the number killed, since we're morbid sort of folks, at least when it comes to insects. Note there is an uncontrolled source of variation in this experiment–the number of insects in each experimental plot to begin with.

[13] The original paper in which the data were reported, however, argues for just such a transformation.

```
Response: count

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     71     409.04
spray  5   310.71        66      98.33 4.979e-65
```

The fit here looks pretty strong. There is something surprising, however, which is the residual deviance is much larger than the degrees of freedom, indicating overdispersion. (Recall the deviance should be approximately χ^2_{df} and thus should be $\approx df$.) We'll come back to this problem in the next lecture.

The fits of this model can be seen using the summary command, along with z test values (using the asymtotic normality of the mle parameter estimates).

```
> summary(spray)

Call:
glm(formula = count ~ spray, family = poisson, data = InsectSprays)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3852  -0.8876  -0.1482   0.6063   2.6922

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67415    0.07581  35.274  < 2e-16 ***
sprayB       0.05588    0.10574   0.528    0.597
sprayC      -1.94018    0.21389  -9.071  < 2e-16 ***
sprayD      -1.08152    0.15065  -7.179 7.03e-13 ***
sprayE      -1.42139    0.17192  -8.268  < 2e-16 ***
sprayF       0.13926    0.10367   1.343    0.179
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 409.041  on 71  degrees of freedom
Residual deviance:  98.329  on 66  degrees of freedom
AIC: 376.59

Number of Fisher Scoring iterations: 5
```

Changing the link function from the default, the log link, to another has some impact on the estimated parameters.

```
> spray2<-glm(count~spray,data=InsectSprays,family=poisson(link=sqrt))
```

```
> summary(spray2)

Call:
glm(formula = count ~ spray, family = poisson(link = sqrt), data = InsectSprays)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3852  -0.8876  -0.1482   0.6063   2.6922

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.8079     0.1443  26.382  < 2e-16 ***
sprayB        0.1079     0.2041   0.529    0.597
sprayC       -2.3645     0.2041 -11.584  < 2e-16 ***
sprayD       -1.5905     0.2041  -7.792  6.6e-15 ***
sprayE       -1.9371     0.2041  -9.490  < 2e-16 ***
sprayF        0.2746     0.2041   1.345    0.179
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 409.041  on 71  degrees of freedom
Residual deviance:  98.329  on 66  degrees of freedom
AIC: 376.59

Number of Fisher Scoring iterations: 4
```

Notably, however, the overall appearance of the estimates is close. The estimated responses, likewise, are pretty close. The maximum difference between the predicted values for the two observations is about 1.2.

```
> max(abs(predict(spray)-predict(spray2)))
[1] 1.269072
```

This is somewhat comforting, since it means predictions from the two models are pretty close, so the choice of link functions has little effect on the model.

Lecture 14

Generalized linear models 2

14.1 Main ideas

- General form of the glm
  - Other glms
  - Fitting a glm
- Diagnostics
  - Overdispersion
  - Deviance residuals
  - Brief summary of other diagnostics

14.2 General form of a glm

Although count regression is perhaps the most commonly encountered problem in ecology which indicates a glm approach, the glm can be used in a much wider set of circumstances. The response can be modeled using any of wide collection of distributions (any exponential family, defined below, and even a few distributions which are not quite exponential families). The link function, likewise, need only satisfy some basic conditions to be used.

14.2.1 Exponential families

Definition 14.2.1 (Exponential family)**.** *A family of probability density or probability mass functions $f(y|\psi)$, dependent on the parameter(s) ψ is an exponential family if f can be written in the form*

$$f(y|\psi) = s(y)t(\psi)\exp(a(y)b(\psi)) \tag{14.1}$$

For glms, it is convenient to consider ψ to have two parts: a canonical parameter θ and a dispersion parameter ϕ.[1] *Then write the exponential family in the form*

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{14.2}$$

This second form may seem a bit peculiar, since both ϕ and θ are functions of the parameters ψ (which is potentially just a single parameter). However this form has some nice properties.[2]

Theorem 14.2.1.

$$EY = \frac{d}{d\theta}b(\theta) \tag{14.3}$$

$$VarY = \frac{d^2}{d\theta^2}b(\theta)a(\phi) \tag{14.4}$$

14.2.2 Requirements of a glm

Glms, like all statistical models in this course, are functions of the predictors which yield the response. The general form of a glm is to model the response Y as a random variable which

- is an exponential family,
- $g(\mathrm{E}Y) = X\beta$ for a smooth, invertible link function g

14.2.3 Other exponential families

There are a lot of exponential families which are used in glms, beyond the two which have already come up. In fact, it is possible to think of the normal distribution (that is, ordinary linear regression) as a generalized linear model–the normal distribution is an exponential family.

Normal model

The term "generalized" linear model seems to indicate that the model is a generalization of the commonly used normal (Gaussian) model used for ordinary regression. This is, in fact the case. Note that the predictors $X\beta$ are the expected value of the response Y, fulfilling the first part of a glm. The normal distribution is an exponential family, and the link function is most often the identity. (Note that a log or power transformation of the response can also be realized as a specific choice of link function.)

Thus it would be possible to consider ordinary linear regression to be a special case of generalized linear regression. The `glm()` command even fits a normal (called `Gaussian` in `R`) model, although in general the estimation using `lm()` works much faster and is more accurate. Never-the-less, all of the ordinary linear model development can be recast as a special case of generalized linear models.

[1] This follows the notation of Faraway ch 6, and ...

[2] This is cheating a bit, since the first version has nice properties too.

Gamma distribution

The gamma distribution is a continuous, two parameter exponential family of distributions. The support of a gamma random variable is the positive reals. It is a generalization of an exponential distribution, which is also defined below (although the formal definitions may be too technical for some, and can be skipped here).

Definition 14.2.2 (Exponential distribution)**.** *In the definition of a Poisson random variable (see definition 0.3.17), a Poisson process was defined. While the Poisson random variable can be realized as the count of events over a specified period of time, an exponential random variable is the distribution of times between events. An exponential random variable has the pdf, depending on the parameter* λ

$$f_{exp}(y|\lambda) = \frac{1}{\lambda}\exp\left(-\frac{y}{\lambda}\right) \tag{14.5}$$

The exponential distribution is commonly used in waiting times for memory-free processes (where the events prior to a time do not influence the subsequent events).

Some sources define a gamma random variable to be the sum of α independent exponential random variables. This is a good motivation, however the definition below generalizes it even further to cases where α is not an integer.

Definition 14.2.3 (Gamma distribution)**.** *A gamma random variable, defined on the support* $y \geq 0$*, with parameters* α *and* β *has the pdf*

$$f_{Gamma}(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-\frac{y}{\beta}} \tag{14.6}$$

where the gamma function $\Gamma(z)$ *(a generalization of the factorial) is defined*

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt \tag{14.7}$$

Using this parameterization[3], the expected value of a gamma random variable is $\alpha\beta$ and the variance is $\alpha\beta^2$. It is convenient to parameterize the model using $\mu = \frac{1}{\alpha\beta}$ and $\phi = \frac{1}{\alpha}$.

There are three possible link functions pre-set in `R` to use with the Gamma glm, summarized in the table 14.1.

Further extensions

In addition to a few other exponential families such as the inverse-Gaussian or multinomial,[4] glms get extended in a few other ways. The binomial and Poisson models may, in some cases, fail to fully capture the behavior observed; one fix is to introduce two-parameter versions of these distributions, termed quasibinomial and quasipoisson, to better account for observed dispersion (these are touched on briefly below). These generally are in the form of quasilikelihood, which again is outside of the course content. Glms have also been developed for distributions which are not quite exponential, such as the negative binomial and the Weibull.

[3] There are, unfortunately, many parameterizations of the exponential distribution and gamma distribution. This one follows Casella and Berger, but is different from Faraway.

[4] Not covered here for brevity, see Faraway for more information

Distribution	Link	Description
Normal	Identity	$\mu = X\beta$, equivalent to ordinary least squares.
	Log	$\log(\mu) = X\beta$
	Inverse	$\mu^{-1} = X\beta$
Poisson	Log	$\log(\lambda) = X\beta$
	Square root	$\sqrt{(}\lambda) = X\beta$
Binomial	Logit	$\log\left(\frac{p}{1-p}\right) = X\beta$
	Probit	$\Phi^{-1}(p) = X\beta$
	Cauchit	$\tan\left(\pi(p - \frac{1}{2})\right) = X\beta$
	C. Log-Log	$\log(-\log(1-p)) = X\beta$
Gamma	Inverse	$\mu^{-1} = X\beta$
	Identity	$\mu = X\beta$
	Log	$\log\mu = X\beta$

Table 14.1: Link function options in `R` for common glm families.

14.3 Fitting the glm

The first goal, as with previous models, is to first estimate the parameters. For glms, like ordinary least squares, the free parameters to be fit are in the vector β. The estimates of β are just the usual maximum likelihood estimates. Fortunately, the numerical method for fitting the parameters does not have to rely on a general optimization algorithm such as Nelder-Mead or BFGS. By taking advantage of the structure of the model, much more efficient methods have been developed.

The fitting routine for glms reduces to iteratively fitting and re-weighting least squares. In ordinary linear models, observations Y are modeled as $N(X\beta, \sigma I_n)$ where I_n is an $n \times n$ identity matrix. A nice extension of this is to change I_n to a diagonal matrix W.[5] The reciprocal of the square root of the diagonal elements of W are known as the "weights" (by convention).

The glm fitting procedure maximizes the likelihood by iteratively fitting and re-weighting least-squares approximations. The full details of this procedure can be found in Dobson (2002) section 4.3.[6]

14.4 Diagnostics

14.4.1 Overdispersion

Overdispersion is a common problem. Broadly, a model fit is overdispersed when it has a greater variance than is expected. Note that in two parameter models, it is generally possible to estimate dispersion and mean separately. In one parameter models, such as the binomial or Poisson, however, the dispersion is fixed once the single parameter is estimated.

[5] In fact, it could be generalized to an arbitrary positive definite matrix Σ, however that is not immediately relevant at the moment.

[6] Ask me if interested.

Overdispersion (or underdispersion–less variance than would be expected) can be detected for one parameter models (which do not estimate a dispersion parameter) using the asymptotic distribution of the residual deviance (definition 13.3.2).[7] The residual deviance should be χ^2_{n-p} distributed (asymptotically); large values of the residual deviance are an indicator that there may be a problem of overdispersion. Since the expected value of a χ^2_z is z, the residual deviance should be approximately $n-p$, the residual degrees of freedom. Values much higher than $n-p$ indicate overdispersion.

Several possible causes exist for overdispersion (and this list is not comprehensive):

- Inaccurate specification of the model (that is, the model is really not Poisson or binomial, etc.)
- Outliers or unusual observations of some type
- Correlated data
- Poor asymptotic convergence to a χ^2 distribution

This course will deal primarily with the first and second of these. Correlated data may be modeled and asymptotic convergence assessed,[8] although doing so is outside the scope of this course.

Correcting for overdispersion, at least where unusual observations can not be identified (see below), can take the form of making a second estimate of the dispersion parameter. Faraway recommends the estimator

$$\hat{\phi} = \frac{1}{n-p} X^2 \tag{14.8}$$

where X^2 is Pearson's goodness of fit statistic (see definition 13.3.1). The model fit can then be assessed using an F test, via equation 13.11. Formally this is part of quasilikelihood estimation (which is a more general process).

For this reason, Dr. Conquest's rule of thumb for Poisson and binomial glms was to use a χ^2 tests for model fit (equation 13.11) for most problems, and an F test (as described) for overdispersed data.

14.4.2 Residuals

In work with ordinary linear models there have been several kinds of residuals used. Raw residuals were the starting point of linear regression, and were modified to form standardized residuals and studentized residuals. Standardized residuals can be used in generalized linear models, but get some new terminology, see table 14.2.

Standardized residuals are useful for glms, since the variance of a non-normal glm is probably not constant. Standardized residuals are a reasonable way to compare the model fit at different points.

Deviance residuals, on the other hand, are perhaps a more natural extension of the idea of a residual. Recall that the deviance of a model was defined (equation 13.9) as negative twice the

[7] "Residual" deviance meaning the deviance associated with the full model, not the deviance associated with a model with fewer parameters.

[8] Well, honestly I don't know of a method of assessing whether an asymptotic approximation should hold, at least in general, other than Monte Carlo simulation. Maybe there is a literature on it which I am unaware of.

OLS term	glm term	Equation
Raw residual	Response residual	$y - \hat{y}$
Standardized residual	Pearson residual	$\frac{y-\hat{y}}{\sqrt{\text{Var}(\hat{y})}}$
Standardized residual†	Deviance residual	$\text{sign}(y-\hat{y})\sqrt{d_i}$
Studentized residual	Studentized residual†	$\begin{cases} \frac{y_i-\hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} & \text{ols} \\ \frac{r_D}{\sqrt{\hat{\phi}(1-h_{ii})}} & \text{glm} \end{cases}$

Table 14.2: Forms of residuals in ordinary linear models and generalized linear models. Here r_D is the residual deviance, d_i is the deviance for observation i, and h_{ii} is the ith diagonal element of the hat matrix (see section 14.4.3 for discussion of the hat matrix in a glm context). †Note that the deviance residual reduces to the Pearson residual (and thus is the same as the standardized residual) for the normal (Gaussian) glm.

log of the likelihood ratio of the saturated model to the model being fit. Since observations Y_i are independent of each other, the joint likelihood of all of them is the product of each likelihood. Using the fact that a log of a product is a sum of logs, the deviance can be written as $\sum d_i$ where there is a deviance for each observation (just the log-likelihood of that observation under the saturated model, minus the log-likelihood of the same observation under the model being fit).

To make the residuals comparable to Pearson residuals, the square root is taken and a sign to indicate the direction of the deviance is multiplied. By examination of the residuals themselves, it may be possible to identify outliers. This would work, provided the point does not also have a large influence (see section 14.4.3).

Faraway recommends a half-normal plot (using the function `halfnorm()` to identify usual observations. A half normal plot compares the ranked residuals to the lower half of a normal density, not particularly because the residuals should be normal. Rather the half normal plot should be smooth, and it is easier to identify what points are unusual.[9]

14.4.3 Other diagnostics

For ordinary linear models, several additional measures to detect unusual observations were developed: leverage, studentized residuals, DFFITS, and Cook's distance. These can all be used for glms as well. Note, however, that all of these rely on a hat matrix, which does not generally exist for glms. To get around this, recall from section 14.3 on fitting glms that iteratively reweighted least squares is used. Thus the final reweighting gives a normal approximation of the model with weighing matrix W. This can be used to develop an analogous hat matrix to the one used in ordinary linear regression.

Faraway gives the following analogue for the hat matrix $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$ for use with glms.[10] Using this formula, it is possible to develop Cook's distance studentized residuals.

[9]This is done for ordinary least squares as well, however the plots are rather obscure to read and not very sensitive, in my view. Because there is no reference (unlike the qqplot, where data should form a rough line), the half norm is difficult to read. Both Faraway books discuss it in more detail than here.

[10]He does not, unfortunately, include a derivation, nor do McCaullagh and Nelder who use the same formula. I have not been able to replicate these results. Thus explaining this formula is left as an extra credit problem.

Figure 14.1: The Brussels sprout, *Brassica oleracea*. Photo by André Karwath (cc-sa-2.5).

14.5 Examples

Name	section	description
Brussels Sprouts	14.5.1	Equivalence of a binomial glm and χ^2 test.
Rodent Capture	??	Unusual observations and overdispersion in a Poisson regression.

14.5.1 Brussels sprouts

Opinions on the Brussels sprout (see figure 14.1) vary widely. If vegetables can be controversial, Brussels sprouts would undoubtedly be among the most devisive. The variability of preferences for Brussel sprouts has been examined by John Trinkaus, an eternal source of interesting research, and Karen Dennis.[11]

They report on a survey done of 442 business students about their preferences for Brussels sprouts. Although the original data used five categories of response (from "very repulsive" to "especially delicious"), since this course only deals with univariate response the data will be grouped together. The "repulsive" to "indifferent" categories will be lumped into one, and the "delicious" categories will be the other. The predictor here will be sex.

Thus a 2×2 contingency table is used. Early in the class this sort of data was dealt with using a χ^2 test, see section 2.5. It is also accessible, however, to a generalized linear model approach using a binomial error structure. These approaches, it turns out, will yield the same answer.

	women	men
Repulsive	247	160
Delicious	19	16

Rather than reading in the data from online, it is perhaps easier just to type it in. (The dataset is online, separated into the five categories instead of two, in `brussels.txt`.)

[11] Trinkaus J. and K. Dennis. 1991. "Taste preferences for Brussels sprouts: an informal look." *Psychological reports* 69: 1165-1166.

```
> bs<-matrix(c(247,17,160,16),2,2)
```

The χ^2 test is also a straightforward command to run. Since it will be compared later to a glm, Yate's correction, which normally occurs for 2×2 tables by default on the χ^2 test, is turned off.

```
> chisq.test(bs,correct=F)

        Pearson's Chi-squared test

data:  bs
X-squared = 1.0702, df = 1, p-value = 0.3009
```

Now try a glm approach. The data will have to be reorganized to fit the requirements.

```
> res<-t(bs)
> x<-factor(c("women","men"))
```

Now the binomial glm can be run using `res` as the response and `x` as the predictor.

```
> bs.glm<-glm(res~x,family=binomial)
> anova(bs.glm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: res

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     1    1.05239
x     1  1.05239         0  3.197e-14   0.30496
```

Note that the glm gives the same significance to the use of x as a predictor as the χ^2. The same information can even be read off a summary table, using a Z test.

```
> summary(bs.glm)

Call:
glm(formula = res ~ x, family = binomial)

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.3026     0.2622   8.782   <2e-16 ***
xwomen        0.3736     0.3628   1.030    0.303
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.0524e+00  on 1  degrees of freedom
Residual deviance: 3.1974e-14  on 0  degrees of freedom
AIC: 13.14

Number of Fisher Scoring iterations: 3
```

(There are minor discrepancies in these results, which I believe stem from the effects of the numerical fitting routines used, and the different approaches to the tests taken.) All of these results indicate there are no significant differences between men and women in their preferences for Brussels sprouts.

14.5.2 Rodent captures

Biogeographic data has come up several times previously. Here the data are from rodent captures in a variety of canyons in California. Predictors are many, so only a few will be examined carefully.[12]

The response, the number of rodent captures in each area, is modeled as Poisson. Predictors in this model can be any of a number of variables, but for now just consider area of the site, distance to the nearest large canyon, and age since the canyon was isolated. These variables are shown below (three sites with missing ages, intended to be used as controls, were excluded).

Captures	Area	Age (yrs)	Distance
7	25.000	50	2100
160	84.100	20	914
50	53.800	34	1676
102	51.800	34	243
93	25.600	16	822
119	32.100	14	121
0	9.700	79	1554
0	8.700	58	1219
0	8.500	36	2865
0	8.400	31	670
0	8.100	74	365
29	7.600	11	550
28	7.500	18	40
0	6.400	56	304
0	6.100	37	2386
0	6.000	23	228
0	5.100	22	662
106	4.800	8	61
1	4.300	86	1767
28	3.900	6	1000
0	3.600	20	609
0	3.500	77	335
12	1.300	2	91
0	1.100	32	883
0	0.410	77	487

[12] I got these data online, there is a reference in the data file for the original paper.

This high zero-count is a frequent occurrence in ecological data, and several methods have been developed to deal specifically with it. However the approach here is a bit different. First, import the data and fit the model.

```
> rod<-read.table(
    file="http://students.washington.edu/nesse/qerm514/data/rodents.txt",
    header=T)
> rod.glm<-glm(Captures~Area+DistX*Age,data=rod,family=poisson)
```

Looking at the analysis of deviance table, even without a test, immediately indicates a problem.

```
> anova(rod.glm)
Analysis of Deviance Table

Model: poisson, link: log

Response: Captures

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev
NULL                    24    1613.22
Area   1   672.31        23     940.91
DistX  1   261.26        22     679.65
Age    1   262.55        21     417.10
```

Note that the residual deviance and residual degrees of freedom are an order of magnitude different. This is a strong indication of problems, particularly with overdispersion. A check of Cook's distance does indicate a particularly unusual observation: point 2. See figure 14.2.

```
> plot(rod.glm,which=4)
```

Nothing about point two in the other predictors makes it particularly suspect, however. One option would be to throw out the data point, or investigate it further. Removing the point and checking the anodev indicates there is still strong indications of overdispersion.

```
> cap<-rod$Captures[-2]
> ar<-rod$Area[-2]
> dis<-rod$DistX[-2]
> ag<-rod$Age[-2]
> rod.glm2<-glm(cap~ar+dis+ag,family=poisson)
> anova(rod.glm2)
Analysis of Deviance Table

Model: poisson, link: log

Response: cap
```

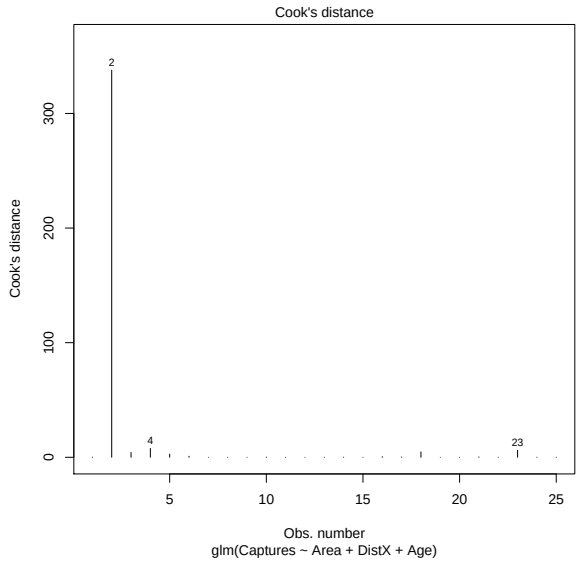


Figure 14.2: A plot of Cook's distance for the rodent capture data. The plot indicates point two has a very large Cook's distance, indicating the observation is, in some way unusual.

```
Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    23    1306.47
ar    1   402.93         22     903.54
dis   1   248.74         21     654.80
ag    1   407.92         20     246.88
```

The conclusion here is it may be better to keep the point and use tests which allow for overdispersion. To set up these tests, first calculate the dispersion parameter using Pearson's residuals.

```
> resids<-residuals(rod.glm,type="pearson")
> disp<-sum(resids^2)/21
```

The 21 here is the residual degrees of freedom for the model. Now the direct calculation of the F test is straightforward. Recall the model deviance associated with age, for instance, was 262.55 with one degree of freedom. Thus the F test associated with that predictor, taking into account overdispersion, is calculated using the `pf()` function.

```
> 262.55/disp
[1] 15.1914
> 1-pf(15.1914,1,21)
[1] 0.0008296375
```

The p value for such a test is 0.000829. A faster way to go, if you're fitting a model with overdispersion, is to fit the `quasipoisson` family rather than Poisson.

```
> anova(rod.glm3,test="F")
Analysis of Deviance Table

Model: quasipoisson, link: log

Response: Captures

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                     24    1613.22
Area   1   672.31        23     940.91 38.900 3.467e-06 ***
DistX  1   261.26        22     679.65 15.117 0.0008488 ***
Age    1   262.55        21     417.10 15.192 0.0008296 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

A check of the diagnostic plots indicates a reasonable model fit for this set up, although notably point two still has a high Cook's distance, albeit much reduced.14.3

Recall this problem of overdispersion came up before, in the Insect Spray data, section 13.4.4.

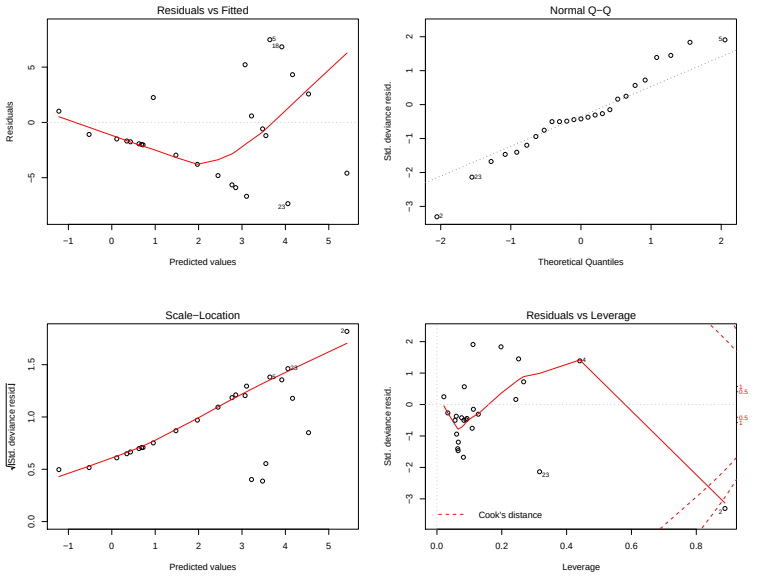


Figure 14.3: Diagnostic plots for the rodent data using the overdispersion. Cook's distance is still high for point 2.

Lecture 15

Generalized linear models, practice

15.1 Testing the χ^2 test for binary data

A situation with binary responses (where the response is either value or another) is a fairly common occurrence. If the predictors are all discrete, responses can be grouped into collections with identical predictors and the usual binomial glm can be applied. Data of this form will be examined later in this lecture. However, when the predictor is continuous, such a grouping can not be reasonably done.

Note that binary response can be modeled as a special case of a binomial random variable, where all the $n_i = 1$. It is similar to a count regression. But what kind of asymptotic results hold; can a χ^2 approximation, or a normal approximation for the parameters be used? Both Faraway and Dobson suggest there could be problems with the asymptotic result. This example uses Monte Carlo sampling to test how good the χ^2 approximation is for binary response data (with a continuous predictor).

The model being considered is a simple binomial glm: each response y_i is a binomial with the same p value, which is not dependent on x_i. Under this model, how often does each test reject the null hypothesis that x in fact has no effect? There are two parameters which might affect the rejection rate: the actual probability of getting a success response (a response of 1) and the number of replicates.

There are several methods which might be used to determine if a model is significant which we might consider. Two such methods are the χ^2 test for deviance (the difference in deviance between the modeling being fit and the null model), or a normal approximation to the fit parameter β_1. The first question of concern is rejection under the null hypothesis. That is, when the data are generated from a binomial with a probability p of success equal for all observations (thus does not depend on a randomly generated predictor), how often does the predictor come up as significant.

In this experiment, 10,000 replicates of the same experiment are performed a fixed number of observations (10,000 replicates with two observations, through 10,000 replicates with 50 observations each). Each replicate has a randomly chosen p (uniformly between 0 and 1).

This simulation (which took several hours to run) is shown below.

```
ChiSqRes<-NULL
NormRes<-NULL
```

```
for(zz in 2:50){
  pres2<-NULL
  nres2<-NULL
  for(tt in 1:10000){
    x<-rnorm(zz,0,5)
    y<-rbinom(zz,1,runif(1,0,1))
    binary.glm<-glm(y~x,family=binomial(link=logit))
    pres2[tt]<-anova(binary.glm,test="Chisq")$P[2]
    nres2[tt]<-summary(binary.glm)$coef[2,4]
   }
  ChiSqRes[zz]<-length(pres2[pres2<.05])/1000
  NormRes[zz]<-length(nres2[nres2<.05])/1000
}
```

The results are then plotted in figure 15.1.

```
plot(2:50,NormRes[2:50]/10,type='l',
  main="Normal coefficient approximation",
  xlab="Number of Observations",
  ylab="Actual type I error rate at alpha = 0.05")
plot(2:50,ChiSqRes[2:50]/10,type='l',
  main="Chi square deviance test",
  xlab="Number of Observations",
  ylab="Actual type I error rate at alpha = 0.05")
```

Both of the approaches seem to have problems for few observations. The χ^2 test on the difference of deviance consistently rejects the null hypothesis more often than it should, which is perhaps the worst problem. This effect is smaller as the number of observations increases, but is somewhat troubling. The normal approximation for the coefficient does the reverse, and underestimates the error rate (which for philosophical reasons is preferred). However for small sample sizes, it virtually never rejects the null, which lead to questions of the power of the test.

In fact, for many binary response data situations, the χ^2 test is not very good. Another approach (which has not been presented in this course) is the Hosmer-Lemeshow statistic. The data are grouped into g groups by similarities in probability, and a χ^2 test is done on those data. There are other approaches as well.[1]

15.2 Aviation deaths

Probably this is the most morbid example in the class, this example looks at the rate of pilot deaths as a function of age. The data are from Australia, and span the better part of the 1990s.

[1] And to be honest, they are new to me. Researching one or more is an extra credit problem.

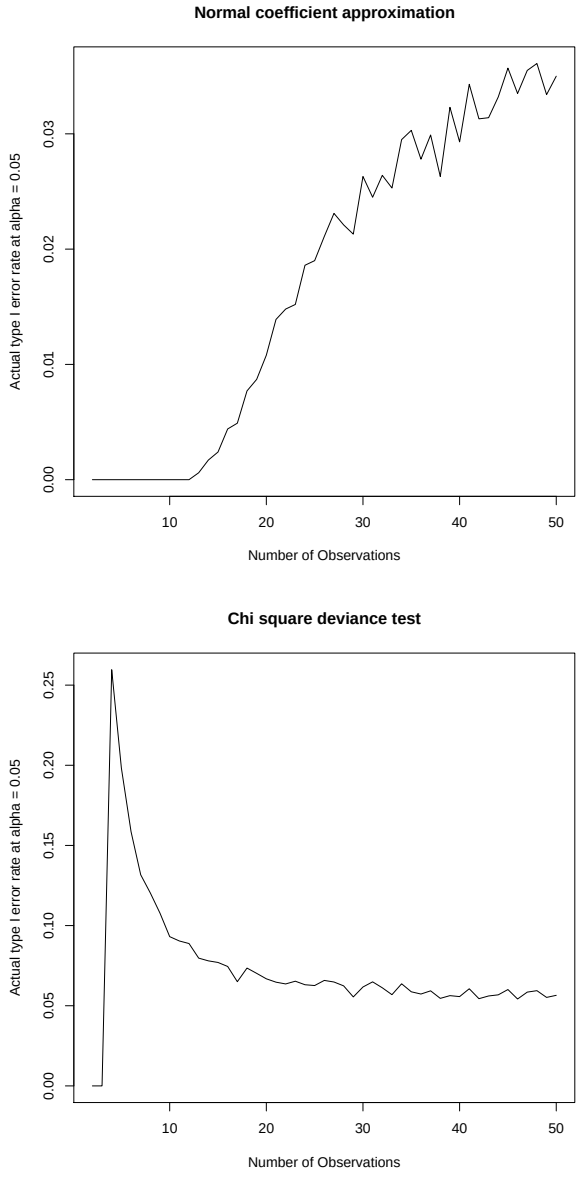


Figure 15.1: Type I error rates for different numbers of replicates, for normal approximation of coefficient (top) and χ^2 test of deviance (bottom). The indicated rate is how many, of 10,000 replicate for each number of observation s (2 through 50) a test at the 0.05 significance level would find a significant result.

Year	Numbers	Deaths	Age	Year	Numbers	Deaths	Age
1992	546	1	0-19	1996	4374	0	0-19
1992	3141	4	20-29	1996	9456	3	20-29
1992	5875	4	30-39	1996	9157	4	30-39
1992	12731	8	40-49	1996	9659	4	40-49
1992	16230	3	50-59	1996	5531	4	50-59
1992	2175	3	60-69	1996	2326	4	60-69
1992	418	0	70-79	1996	630	0	70-79
1992	24	0	80	1996	33	0	80
1993	6278	0	0-19	1997	3009	0	0-19
1993	13026	1	20-29	1997	6849	2	20-29
1993	12380	2	30-39	1997	7098	2	30-39
1993	10449	4	40-49	1997	7934	3	40-49
1993	5452	7	50-59	1997	5533	4	50-59
1993	2534	2	60-69	1997	2176	2	60-69
1993	532	0	70-79	1997	713	3	70-79
1993	27	0	80	1997	37	0	80
1994	5166	0	0-19	1998	2316	0	0-19
1994	11204	4	20-29	1998	5035	1	20-29
1994	10774	1	30-39	1998	5172	4	30-39
1994	10539	4	40-49	1998	7846	8	40-49
1994	7037	4	50-59	1998	5328	8	50-59
1994	2448	4	60-69	1998	2931	2	60-69
1994	595	3	70-79	1998	718	1	70-79
1994	27	0	80	1998	35	0	80
1995	3711	1	0-19	1999	1753	1	0-19
1995	8086	4	20-29	1999	3265	2	20-29
1995	8078	2	30-39	1999	3286	2	30-39
1995	8740	4	40-49	1999	5371	3	40-49
1995	7762	11	50-59	1999	4881	4	50-59
1995	2253	3	60-69	1999	1962	4	60-69
1995	656	1	70-79	1999	728	1	70-79
1995	35	0	80	1999	43	0	80

There are naturally several models which could be fit here. Year and age could both be predictors, and the model fit would be akin to a contingency table, with one observation per combination of levels. However there is little *a priori* reason to suspect changes in the fatality rates from year to year. As such, years will be treated as independent observations of the same variables, improving the ability to estimate.

There are several hypotheses which may be testable with these data. It might be reasonable to expect pilot deaths to decrease as a function of age, due to increasing experience of the pilots. Pilots in most countries, including Australia[2] must pass a health exam to fly. However undetected health problems may also increase with age which interfere with the ability to fly.

It is natural to code the age as an ordered predictor.

```
> av<-read.table(
    file="http://students.washington.edu/nesse/qerm514/data/avdeath.txt",
    header=T)
> attach(av)
> av.bin<-glm(Deaths/Numbers~as.ordered(Age),weights=Numbers,family=binomial)
> anova(av.bin, test="Chisq")
Analysis of Deviance Table
```

[2]From what I could find online

```
Model: binomial, link: logit

Response: Deaths/Numbers

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                              63    118.139
as.ordered(Age)  7   52.091        56     66.048 5.602e-09
```

These data do indicate age is a significant predictor of the mortality rate among pilots. Note that the anodev table would not change if the coding were altered, just as was the case for anova tables. There is an indication of a small degree of overdispersion, as the residual deviance is about 10 points higher than the degrees of freedom. It is not enough overdispersion, however, to be concerning.

To look for trends in the data, the summary table (and z tests) are used.

```
> summary(av.bin)

Call:
glm(formula = Deaths/Numbers ~ as.ordered(Age), family = binomial,
    weights = Numbers)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.2657583  -0.6445941  -0.0002223   0.5905112   2.1307223

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.2448   188.6614  -0.049    0.961
as.ordered(Age).L  -5.5594   815.1104  -0.007    0.995
as.ordered(Age).Q  -8.4034   815.1104  -0.010    0.992
as.ordered(Age).C  -6.3610   650.2332  -0.010    0.992
as.ordered(Age)^4  -4.6400   425.6776  -0.011    0.991
as.ordered(Age)^5  -2.0069   226.0711  -0.009    0.993
as.ordered(Age)^6  -1.0578    92.8907  -0.011    0.991
as.ordered(Age)^7  -0.1711    25.7638  -0.007    0.995

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.139  on 63  degrees of freedom
Residual deviance:  66.048  on 56  degrees of freedom
AIC: 222.85

Number of Fisher Scoring iterations: 17
```

Interestingly here, there are no significant trends of any order. (Of course, this does *not* contradict the significance under the χ^2 test in the analysis of deviance–it may be there are significant differences but in no particular order.) This may be taken as falsification of the hypothesis, although that would be a dull example if just left there. There is something more going on here. Take a look at the summary table under the (default) treatment coding.

```
> av.bin<-glm(Deaths/Numbers~Age,weights=Numbers,family=binomial)
> summary(av.bin)

Call:
glm(formula = Deaths/Numbers ~ Age, family = binomial, weights = Numbers)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.2657583  -0.6445941  -0.0002223   0.5905112   2.1307223

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.1105     0.5774 -15.779  < 2e-16 ***
Age20-29       1.1523     0.6173   1.867  0.06194 .
Age30-39       1.1234     0.6173   1.820  0.06876 .
Age40-49       1.5467     0.5997   2.579  0.00991 **
Age50-59       1.9540     0.5963   3.277  0.00105 **
Age60-69       2.4480     0.6124   3.997 6.41e-05 ***
Age70-79       2.7944     0.6668   4.190 2.78e-05 ***
Age80        -12.0928  1509.2912  -0.008  0.99361
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.139  on 63  degrees of freedom
Residual deviance:  66.048  on 56  degrees of freedom
AIC: 222.85

Number of Fisher Scoring iterations: 17
```

Recall under this coding, each fit parameter other than the intercept can be interpreted as the difference from the intercept. These data *do* show an increasing trend in the probability of mortality with age. The intercept is fit parameter for the lowest group, 0-19 ages. The next group, 20-29, has a mortality probability that is a bit higher than them, but really marginally significant. Each increasing age has a slightly higher probability of dying (except for the 30-39, which is almost identical to 20-29), up to the last group. The 80+ age group is much lower than any of the other groups, but this difference is not significant. This should come as no surprise, since there were no deaths in the 80+ group in any year. There were also almost no pilots (an order of magnitude less than any other class) in the 80+ age class.

Figure 15.2: Salmonella bacteria, source: National Institutes of Health

15.3 Food Poisoning

Keeping with the upbeat theme of this lecture, this example discusses food poisoning at a company picnic.[3] Attendees at the picnic were polled with three response variables: Sick or not, eat potato salad or not, and eating crab or not. The responses are shown below as sick/notsick

	Crab	No Crab
Potato Salad	120/80	22/24
No Potato Salad	4/31	0/23

These data can be analyzed using a glm (although just looking at the data should indicate the most likely culprit). The data are online, but it is probably just as fast to type in the data manually.

```
res<-matrix(c(120,4,22,0,80,31,24,23),4,2)
crab<-as.factor(c("crab","crab","nocrab","nocrab"))
potato<-as.factor(c("potato","nopotato","potato","nopotato"))
fp.glm<-glm(res~crab+potato,family=binomial)
```

A check of the anova table here confirms what might be evident from reading the table: potato salad was the culprit.

```
> anova(fp.glm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: res
```

[3]From what I can tell this was real data from an insurance company picnic. I suppose it would have to be an insurance company to have the first response to an outbreak of food poisoning is to distribute survey questionnaires.

```
Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       3     63.196
crab    1    9.513         2     53.683     0.002
potato  1   50.940         1      2.743 9.524e-13
```

The chi^2 test was used and there is no real evidence of overdispersion. An oddity here, however, is that the crab is also a significant predictor. It seems unlikely that two dishes at the same picnic would be dangerous. One thing which might be going on here is correlated predictors. Knowing a person ate crab is a reasonable indicator of potato salad consumption.[4] Letting that term enter second in the model is a reasonable approach to looking at this.

```
> fp.glm<-glm(res~potato+crab,family=binomial)
> anova(fp.glm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: res

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       3     63.196
potato  1   56.714         2      6.482  5.04e-14
crab    1    3.739         1      2.743     0.053
```

The significance of crab has dropped dramatically. We're left with a mildly significant predictor, which may either require no explanation, or something which could be due to food sharing, misreporting, reuse of utensils, etc. A look at the summary table shows the fit of the parameters.

```
> summary(fp.glm)

Call:
glm(formula = res ~ potato + crab, family = binomial)

Deviance Residuals:
      1        2        3        4
-0.1566   0.6300   0.3203  -1.4895
```

[4] Another instance where this might be even more pronounced, suppose there were exactly two entrees, and each person had exactly one. If one entree was poisonous, both would be good predictors, since knowing you had the non-toxic one is the same as knowing you didn't have the toxic. In such a case, it would be impossible to determine which entree was poisonous from the analysis of deviance alone. It would be necessary to look at the fit parameters.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.3978     0.5269  -4.551 5.35e-06 ***
potatopotato   2.8259     0.5362   5.271 1.36e-07 ***
crabnocrab    -0.6097     0.3170  -1.923   0.0544 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 63.1957  on 3  degrees of freedom
Residual deviance:  2.7427  on 1  degrees of freedom
AIC: 21.889

Number of Fisher Scoring iterations: 4
```

Potato salad does seem to have a positive effect on the probability of getting sick. Crabs, likewise, have a positive effect on the probability of getting sick (not the factor shown is No Crab), but substantially less than the magnitude of the effect for potato salad.

Lecture 16

Linear mixed effects

16.1 Main ideas

- Random effects
- Blocking and Mixed effects
- Visualizing the data
  - Design plots
  - Trellis plots
- Lme model
- Fitting methods

Computing note: There are two common packages in `R` for linear mixed effects, each of which has different functional calls. The `nlme` package is older, but is on lab computers, while `lme4` is newer and is used in Faraway, but has not yet been installed on the lab computers. These notes will have examples from both packages.

16.2 Random effects and mixed effects

It is not unusual in ecology to encounter hierarchical data. Modeling such data fits some (or all) of the regression coefficients as themselves random variables from another normal distribution. The questions of interest, in such a model, are not the values of the random variable parameters, but of the higher level distribution. The term "random effect" refers to a factor which is modeled as a (normal) random variable, contrasted against "fixed effects" which are simple fit parameters.

Random effects and mixed effects (models which include both random and fixed effects) often arise in situations with correlated observations, something largely avoided in this course. Repeated observations of a person, an experimental patch, a particular animal or plant, etc. often give rise to mixed effects models. The person or patch, for instance, can be thought of as a random sample from

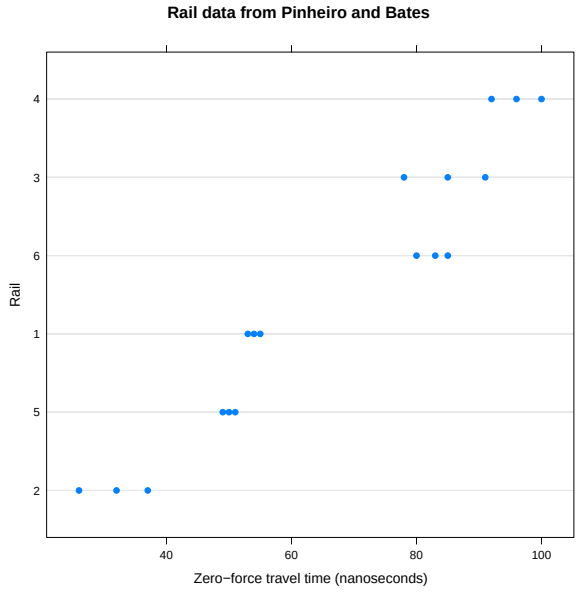


Figure 16.1: A dot plot of the `Rail` data from the `nlme` package. The data shows three measurements on each of six railroad rails of the time for a sound wave to traverse the length of the rail (a measurement of quality). The question of interest is the travel time for the average rail, so a random effects model might be appropriate.

a larger population of interest; it is that population which is of interest, not the specific person[1] or patch. Rather, the questions of concern are the parameters underlying the distribution of people or patches (or animals, or whatever else was modeled in this way).

16.3 Visualizing the data

Data can come in a wide variety of forms, however plotting it can prove challenging. Often there is a default type of plot for the data type in `R`, so often the command `plot()` simply produces an appropriate graph.

16.3.1 Grouped data

Ordinary data frames will produce box plots, which are useful when there are a few groups with multiple observations in each. Many random effects models use grouped data, however, which have many groups and a few observations per group. For instance, the data shown in figure 16.1 is a plot (this one is a method in the `nlme` package) of grouped data, where there are only a few observations per group.

[1] Unless they are a celebrity, in which case we're fascinated for rather mysterious reasons.

The data can be accessed by loading the `nlme` library and plotting the data. Note that the data (see second command below) are grouped data, not an ordinary data frame.

```
> library(nlme)
> Rail
Grouped Data: travel ~ 1 | Rail
   Rail travel
1     1     55
2     1     53
3     1     54
4     2     26
5     2     37
6     2     32
7     3     78
8     3     91
9     3     85
10    4     92
11    4    100
12    4     96
13    5     49
14    5     51
15    5     50
16    6     80
17    6     85
18    6     83
> plot(Rail, main="Rail data from Pinheiro and Bates")
```

Of course it is a reasonable question of how to get data to be grouped. In the `nlme` package, there is a `groupedData()` to transform ordinary data (such as in data frames or vectors) into grouped data. Here `x` will be the data frame of the same data as in `Rail`. To transform it into grouped data as above,

```
> groupedData(travel~ 1|Rail,data=Rail)
```

16.3.2 Blocked experiments

One common use of mixed effects models is blocked experiments. In a blocked experiment, responses are usually the result of block, treatment, and randomness. Often, the specific effects of the block are not of a major concern; instead, the question is likely to center on the treatment on a randomly chosen block. Thus it seems reasonable to model the blocks as random effects.

In most cases of randomized block designs, each block will have all the treatments within it. To compare the effect of the treatments and blocks, a design plot is often used. A design plot is shown in figure 16.2 for each of nine varieties of barley, planted in two different years at six different sites.

```
> plot.design(barley)
```

(Note that the barley data is built into `R`)

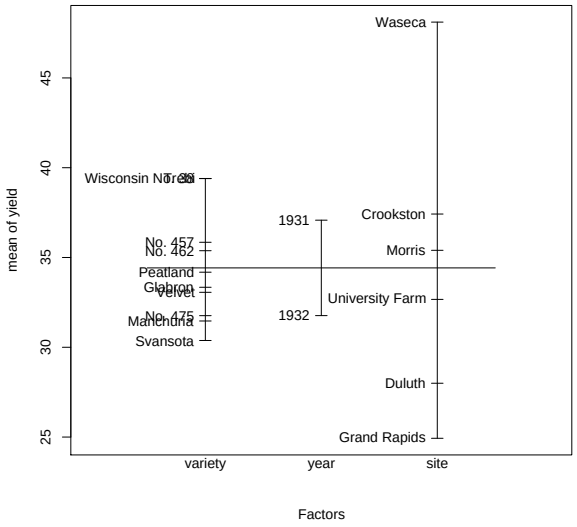


Figure 16.2: The effect of year, block, and grain variety on yield of barely, to demonstrate a design plot.

16.4 Trellis plots

Trellis plots are a large class of plots, which have a separate package (`lattice`) which provides the commands (`lattice` is called by `nlme` and loads automatically). One of the possible uses of trellis plots is to longitudinal data, in fact it is the default method in `R`.

The data are the distances between the pterygomaxillary fissure[2] and the pituitary in children at different ages. Four measurements were taken on each child.

The trellis plot of these data, shown in figure 16.3, plots the repeated observations at each age–8, 10, 12, and 14–for each child (labeled M01 through M16 for boys and F1 through F11 for girls). Note that this plot gives a good indication that, in most children, the measured distance increases with age.

```
> library(nlme)
> plot(Orthodont)
```

16.5 Mixed effects models

The formal structure of a mixed effects model is a bit similar to previous linear models, except that it has two levels of random variables. In addition to an additive normal error, some of the parameters

[2] A gap in the bone between the sphenoid and maxillary bones in the skull.

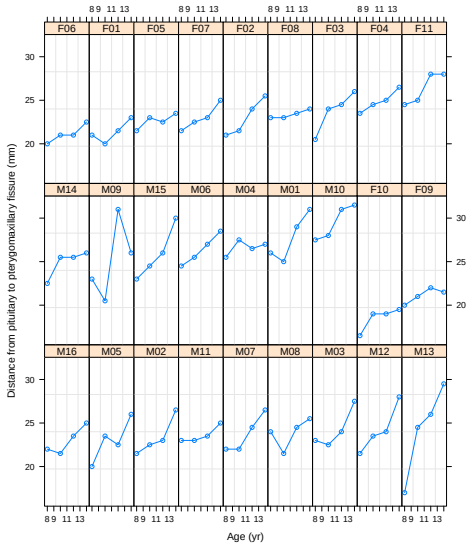


Figure 16.3: Orthodontic data: Distances between the pituitary gland and pterygomaxillary fissure in 27 children (16 boys and 11 girls), at ages 8 though 14.

(which would have otherwise been fit) are included as random variables. Those parameters are modeled to have come from another normal distribution.[3]

Each response i in group j, Y_{ij} is modeled to be

$$Y_{ij} = X_{ij}\beta + W_{ij}b_j + \epsilon_{ij}, \tag{16.1}$$

where the b_j are drawn from a multivariate normal distribution with (positive definite) variance-covariance matrix Ψ. Note that the matrix may have off-diagonal elements.

16.6 Fitting linear mixed effects models

Linear mixed effects models are one of the few models in this course which are not usually fit with maximum likelihood. Although ML is an option, more commonly restricted maximum likelihood (REML) is the default, since it tends to give better estimates. A third option exists for most models: anova fits, which are straight-forward to calculate but may give strange answers.[4]

Recall that there was one other instance where the mle was not used: calculation of the standard deviation (or variance) in the ordinary linear model. In that case, as it is in this case, the mle is biased, so an adjustment is made to the estimate to make it unbiased.

[3]This actually describes only one level above the observation; it is possible to develop a model (along the same lines) which generalizes to multiple levels (groups within groups), but it is not done so here. One reference with a great deal of detail on lme models is Pinheiro and Bates.

[4]In particular, variance estimates can be negative; anova estimates for lmes are not widely used for this reason. As near as I can tell, there is no function in `R` to implement it.

16.7 Examples

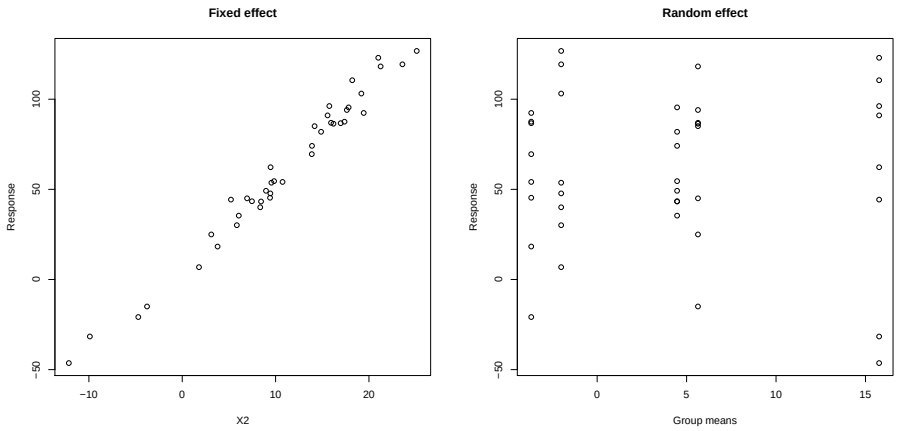
16.7.1 Fake data

To see how the commands work, and to take a look at the generation of the data, this example (i) generates some data which is then fit using (ii) `nlme` and (iii) `lme4` packages. The data being generated are from 5 groups, each with a different mean μ (mean of these means is 3 and a standard dev. of 8), while an unrelated second fixed effect is a continuous random variable. The response is (in reality) $2 + 5 * x_2 + \mu$.

```
> set.seed(1)
> x1<-rep(rnorm(5,3,8),8)
> x2<-rnorm(40,10,10)
> y<-rnorm(40,2+x1+5*x2,3)
>
> ## Fit the model
> summary(lmer(y~x2+(1|xr)))
Linear mixed-effects model fit by REML
Formula: y ~ x2 + (1 | xr)
   AIC   BIC logLik MLdeviance REMLdeviance
 217.3 222.3 -105.6      211.2        211.3
Random effects:
 Groups   Name        Variance Std.Dev.
 xr       (Intercept) 44.5901  6.6776
 Residual              7.4402  2.7277
number of obs: 40, groups: xr, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   5.7668     3.0695    1.88
x2            5.0540     0.0522   96.82

Correlation of Fixed Effects:
   (Intr)
x2 -0.184
```

Before examining the data, note that these data are pretty good.

```
> plot(x1,y, main="Random effect", xlab="Group means",ylab="Response")
> plot(x2,y, main="Fixed effect", xlab="X2",ylab="Response")
```



Now take a look at how well the model did at recovering the information in the model. The estimated standard deviation of the random effect is found to be 6.67, which is not terribly different from the true value of 8 (recall this is estimated using roughly five values in this model). The fixed effects estimated intercept is 5.77. The true value should be the fixed effect intercept (2) plus the mean of the random effect groups (3), so it is 0.77 too high, which is not too bad. Of course, this model will never give information to separate the mean of the random effects groups from the fixed effects intercept (in fact, it is really a nonsensical question). The slope of the fixed effect was 5 by design, and is fit to 5.054 here, so it is a really good estimate.

These data can also be fit using the `nlme` package.

```
> lme(y~x2,random=y~1|xr)
Linear mixed-effects model fit by REML
  Data: NULL
  Log-restricted-likelihood: -105.6326
  Fixed: y ~ x2
(Intercept)          x2
   5.766749    5.054007

Random effects:
 Formula: y ~ 1 | xr
        (Intercept) Residual
StdDev:    6.677582 2.727667

Number of Observations: 40
Number of Groups: 5
```

16.7.2 Pulp

These data come from Faraway, and relate the brightness of paper being manufactured, with plant operator as a predictor.[5] Each of four operators has five measurements taken on his or her work.

[5] I must say, the descriptors of the data, `operator` and `bright` are just begging to have another story made up about the data. Alas, I believe the data are from an actual paper.

The data, in dataset `pulp`, however it has been added online so as to be accessible on lab computers.

```
> pulp<-read.table(file="http://students.washington.edu/nesse/qerm514/data/pulp.txt",
    header=T)
> pulp
   bright operator
1    59.8        a
2    60.0        a
3    60.8        a
4    60.8        a
5    59.8        a
6    59.8        b
7    60.2        b
8    60.4        b
9    59.9        b
10   60.0        b
11   60.7        c
12   60.7        c
13   60.5        c
14   60.9        c
15   60.3        c
16   61.0        d
17   60.8        d
18   60.6        d
19   60.5        d
20   60.5        d
```

The data are plotted in figure 16.4 (using a dot plot discussed in section 16.3.1), which required grouping.

```
> pp<-groupedData(bright~1|operator,data=pulp)
> plot(pp)
```

Note that ordinary anova can be used here.

```
> pul.lm<-lm(bright~operator,data=pulp)
> anova(pul.lm)
Analysis of Variance Table

Response: bright
          Df  Sum Sq Mean Sq F value  Pr(>F)
operator   3 1.34000 0.44667  4.2039 0.02261 *
Residuals 16 1.70000 0.10625
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The model is fairly significant using an F test. It is even possible to look at the estimates of the different operators effectiveness. For ease of interpretation, the coding here will be changed to one mean per group (which then means the F test will be invalid as reported below–the one above is still valid).

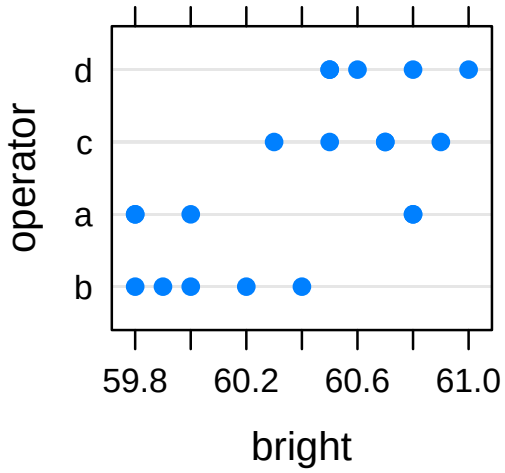


Figure 16.4: Paper brightness for several operators of a paper manufacturing line.

```
> pul.lm<-lm(bright~operator+0,data=pulp)
> summary(pul.lm)

Call:
lm(formula = bright ~ operator + 0, data = pulp)

Residuals:
   Min     1Q Median     3Q    Max
-0.440 -0.195 -0.070  0.175  0.560

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
operatora  60.2400     0.1458   413.2   <2e-16 ***
operatorb  60.0600     0.1458   412.0   <2e-16 ***
operatorc  60.6200     0.1458   415.9   <2e-16 ***
operatord  60.6800     0.1458   416.3   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.326 on 16 degrees of freedom
Multiple R-Squared:     1,      Adjusted R-squared:     1
F-statistic: 1.717e+05 on 4 and 16 DF,  p-value: < 2.2e-16
```

Under this coding, each group is fit with a mean corresponding to a coefficient. Thus the tests are not all that interesting since each group mean is far away from zero (and the F test is invalid as it is not a nested model arrangement). Nevertheless, this analysis might be performed if the question of interest is, for instance, which operator is performing the best.[6] Here, `operatord` appears to be doing the best.

A different question arises if it is assumed that the operators tested actually represent a larger class of operators. In such a case, the researcher might be interested in what the average operator does (and how variable the average operator is). That is a question better suited to mixed effects (or here, just random effects) models.

```
> pul.lme<-lme(bright~1,random=~1|operator,data=pulp)
> summary(pul.lme)
Linear mixed-effects model fit by REML
 Data: pulp
      AIC      BIC  logLik
  24.6262 27.45952 -9.3131

Random effects:
 Formula: ~1 | operator
        (Intercept) Residual
StdDev:   0.2609286  0.32596

Fixed effects: bright ~ 1
            Value Std.Error DF  t-value p-value
(Intercept)  60.4 0.1494437 16 404.1655       0

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-1.4666202 -0.7595239 -0.1243466  0.6280955  1.6012410

Number of Observations: 20
Number of Groups: 4
```

Interpreting this is a bit more complex than for a fixed effects model alone. From the top, the first line indicates that the default restricted maximum likelihood method was used. The next line indicates the data was `pulp`.[7] The next lines, reporting AIC and BIC, will be dealt with in a more general context in a later lecture, although log likelihood shold already be familiar.

The random effects fit is, for the question we are most interested in, the key value. The intercept with a standard deviation of ≈ 0.26 indicates the variability in mean brightness *between workers* (something that was not assessed in the fixed effects model). This is approximately the standard deviation among group means from the fixed effects summary (although it was not reported in the table).

```
> x<-c(60.2400,60.0600,60.6200,60.6800)
```

[6] Well, in fact, I would still not go to this fixed at zero model; instead relying on all pairwise comparisons using contrasts. But since that is not the focus of this example, I'll move on.

[7] In case we forgot, I suppose.

```
> sd(x)
[1] 0.2988868
```

16.7.3 Orthodontic growth

The orthodontic data, shown in figure 16.3, is morphological measurements of boys and girls ages 8 to 12. Each individual is measured four times, and the change as a function of age is a random effect. Since the data are already loaded into R, the grouping is already defined.

```
> lme(Orthodont)
Linear mixed-effects model fit by REML
  Data: Orthodont
  Log-restricted-likelihood: -221.3183
  Fixed: distance ~ age
(Intercept)         age
 16.7611111   0.6601852

Random effects:
 Formula: ~age | Subject
 Structure: General positive-definite
            StdDev    Corr
(Intercept) 2.3270339 (Intr)
age         0.2264276 -0.609
Residual    1.3100399

Number of Observations: 108
Number of Groups: 27
```

Lecture 17

Model selection

17.1 Main ideas

- Forward selection
- Backwards selection
- Problems with inference
- AIC and related approaches
- Cross validation

One of the fundamental questions which comes up in statistical modeling is the topic of model selection. In a very real way, model selection is fundamental to the process of science. Note that the process of model selection is related to, although distinct from, inference techniques.

Model selection, from a statistical point of view, is generally the development of criteria for choosing a model from a collection of possible models. The process, however, is really one which must first start with a scientific step of deciding what models are likely to be reasonable. In an ideal case, it would be nice to identify the model which should match reality from scientific reasoning alone. Failing that, however, some statistical approach can be used to assess model parsimony under the view that a parsimonious model is more likely to be correct.

17.2 Problematic selection criteria

17.2.1 R^2

It might seem reasonable to try something like R^2, the proportion of the sum of squares captured by the model. It can not be reasonably used, however, since the R^2 value always goes up with the addition of new predictors (and does not exist for many nonlinear models). Using it will lead to over fitting–including more parameters than is justified.

A modification of R^2, the adjusted R^2 generally called R_a^2 (see definition 3.5.2) does do some accounting for the number of parameters which enter the model. Thus it does not necessarily go

up. Using it as a model selection criteria does tend toward over fitting still, but less dramatically than the ordinary R^2.

17.2.2 p-values

A routine approach, although one which can be problematic, is to base the model selection on the p value of the predictor. The issue here is that by using p values, the actual meaning of the p value is lost. If only the significant predictors are retained in a model, then it should come as no surprise that the model is significant. The p value for a predictor does take into account the number of degrees of freedom in the model (in the null t distribution for β_i for instance), however it is not included in any kind of integral way. Thus there may be a tendency to include more predictors than is actually necessary.

17.3 AIC

Almost certainly the most common method for assessing model parsimony in ecology is Akaike Information Criteria (AIC) or its derivative measures, AIC_c and BIC. Each is a way of assessing the model fit while penalizing for having more parameters. The question becomes what is the optimal weighting between these two competing interests?

Definition 17.3.1 (AIC)**.** *The AIC is based on the likelihood of the data y in the model with p fit parameters, as a vector $\hat{\theta}$.*

$$AIC = -2\log(f(y|\hat{\theta})) + 2p \tag{17.1}$$

The formula of AIC looks remarkably like Deviance. Akaike suggested that a good criteria would maximize the likelihood (or equivalently, minimize the negative log likelihood). However he proved that this would over fit the model, as judged by information theoretic arguments, so a correction factor of $2p$ had to be added to improve the result.

A correction factor used for small sample sizes is also employed. Corrected AIC, or AIC_c is intended for situations were the number of observations n is small relative to the number of fit parameters. However many people suggest that there is no harm in always using AIC_c.

Definition 17.3.2 (AIC_c)**.** *A correction factor is multiplied on the end of the expression for AIC to correct it for small sample sizes. If there are n observations and p fit parameters, the small sample corrected AIC is*

$$AIC_c = -2\log(f(y|\hat{\theta})) + 2p\left(\frac{n}{n-p-1}\right) \tag{17.2}$$

The third common method in ecology for gauging parsimony is Bayesian Information Criteria (BIC) also called Schwartz's Information Criteria (SIC). The justification for its use was originally Bayesian, however it may be applied in non-Bayesian cases.[1]

Definition 17.3.3 (BIC)**.** *Again, like AIC, BIC is based on the same principle as AIC, with a different adjustment for the number of parameters p. Additionally, however, the number of observations n is also included.*

$$BIC = -2\log(f(y|\hat{\theta})) + p\log(n) \tag{17.3}$$

[1] From what I understand, there is a substantial development of Bayesian model selection techniques beyond this one.

Numerous other formulas exist, each of which balance the fit of the model with the number of estimated parameters. These methods include TIC, Mallow's Cp, and a generalization DIC. When dealing with overdispersed glms, using AIC is not immediately accessible since the overdisperion parameter is not part of the likelihood. However AIC has been modeled for these quasiliklihood situations under the name QAIC.

17.4 Crossvalidation

Another general approach exists to examining parsimony of a model. Akin to the scientific method,[2] models which do a good job of predicting points not used to fit the model should be preferred to those which are poor predictors (even when the fit using all the data is a good one). This suggests a procedure for validating the model: remove one or more points, fit the model to the remaining points and look at how well the missing points are predicted. This technique is generally called "crossvalidation."

Some of the idea of cross validation has already been introduced. The DFFITS technique for finding unusual observations has a similar approach.

17.4.1 PRESS

One of the standard statistics is the Predicted Residual Sum of Squares (PRESS). It is calculated as

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{[-i]i})^2 \tag{17.4}$$

Recall that $\hat{y}_{[-i]i}$ is the predicted observation i using the curve fit to the line not including point i. The choice of using a sum of squares is fairly atheoretic, absent specific knowledge of the distributions involved. For linear ols models, the PRESS statistic is also fairly straightforward to calculate, since the deleted predictions $\hat{y}_{[-i]i}$ can be calculated without refitting the model (although with computing speed available today, for most models and reasonably small datasets, repeatedly refitting a model is feasible).

17.5 Selection process

After deciding on a selection criteria–AIC, PRESS, etc.–there are several methods for applying them. Commonly the approach is to fit all reasonable models, and take the one with the lowest AIC or PRESS. However often this is computationally difficult (and if applied without care, may yield with unreasonable models). Two common approaches to building models have been suggested: forward selection and backward selection. In forward selection, components of the model (such as interaction terms) are added until the selection criteria begins to go up. Backward selection starts with a model with all interaction terms, and progressively drops them out until the selection criteria goes up.

Forward selection and backward selection both generally rely on the view that interaction terms should only be added if the additive terms are included already. Thus the linear (additive) terms

[2] Why does "the" scientific method say we should hypothesize, then test the hypothesis (as opposed to coming up with post hoc hypotheses for the data after the data are gathered)?

are the first ones to enter the model in forward selection, while the highest order interactions should be the first to consider dropping for backwards selection.

17.6 Examples

17.6.1 Problems of data-snooping

This example is intended to highlight the possible *misuse* of model selection criteria. This is not a good example to emulate.

Suppose there are only 15 observations, and a collection of 50 possible predictors. Doing a model selection here can be tricky–the best course of action would probably be to start with deciding, for scientific reasons, which are likely the best predictors. However it may be interesting to see what could go wrong if we blindly plunge ahead with "model selection". First, generating the fake predictors and find the PRESS, AIC, and p value for each of the models $y = \beta_0 + \beta_1 x_i$.

```
> library(MRV)
> set.seed(1)
> y<-rnorm(15,0,5)
> x<-matrix(rnorm(50*15,0,5),15,50)
> p.res<-NULL
> pressres<-NULL
> AICres<-NULL
> for(tt in 1:50){
      model1<-lm(y~x[,tt])
      p.res[tt]<-anova(model1)$Pr[1]
      pressres[tt]<-PRESS(model1)
      AICres[tt]<-AIC(model1)
  }
```

In all three cases, the lowest p-value, PRESS, and AIC belong to predictor 27. In this sense, all three do poorly. However AIC is helpful to spot the meaninglessness of the result. The AIC value for point 27 is only somewhat different from the next lowest value.

```
> sort(AICres)
 [1] 83.41035 91.66256 91.99248 92.53088 93.09925 93.12200 93.50629 93.66146
 [9] 93.82296 93.82627 93.84934 94.17582 94.23872 94.61770 94.87227 94.91328
[17] 95.03735 95.36624 95.48437 95.52937 95.84307 95.85511 95.86221 95.91435
[25] 96.00468 96.05053 96.09158 96.11502 96.14851 96.15695 96.17376 96.18696
[33] 96.20089 96.21664 96.22051 96.23942 96.24719 96.24985 96.25086 96.26195
[41] 96.26213 96.26572 96.26862 96.27785 96.28056 96.31200 96.31370 96.31761
[49] 96.33182 96.34243
```

Nevertheless, no matter the criteria applied we always get a significant model from garbage. This should be seen as a misuse of model selection criteria–it does not do what we would like and indicate the best model independent of the model fit. In fact, comparing these AIC values for these simple models is silly since there is no differences in the number of parameters (it is simply comparing the likelihoods for different sets of predictors).

17.6.2 Marmot trapping

To estimate the abundance of marmots, which may be difficult to trap individually, a proxy measure such as the number of whistles might be used. This was the problem on the midterm. Four possible predictors could be fit, with a response of trapping abundance. Whistles is a good predictor to start with, since it will form the basis for the test. Additional covariates might be considered, however, if the number of whistles might vary significantly in response to other factors. For instance, a region with more predators might have more marmot whistles without necessarily having more marmots.

We might reason that the path through which each of the predictors influences the response falls into two groups: those which predict abundance independently of whistles (alt and lat) and those which influence whistles (thus affecting the ability to predict correctly). Of course, predators might influence the abundance directly as well (lots of predators might exist where there are lots of prey, for instance). This may suggest, however, that predators should be included in the model as well.

Running the AIC and PRESS for these models is not too difficult. The `step()` function could be used for the AIC, however it is not too many models to calculate all possible combinations (including whistles).

```
> marmot<-read.table(
+     file="http://students.washington.edu/nesse/qerm514/data/trap.txt",
+     header=T)
>
> mar.lm<-lm(trap~whistles+predator+alt+lat,data=marmot)
> mar1.lm<-lm(trap~whistles+predator+alt,data=marmot)
> mar2.lm<-lm(trap~whistles+predator+lat,data=marmot)
> mar3.lm<-lm(trap~whistles+alt+lat,data=marmot)
> mar4.lm<-lm(trap~whistles+predator,data=marmot)
> mar5.lm<-lm(trap~whistles+lat,data=marmot)
> mar6.lm<-lm(trap~whistles+alt,data=marmot)
> mar7.lm<-lm(trap~whistles,data=marmot)
> PRESS(mar.lm)
[1] 46.20444
> AIC(mar.lm)
[1] 56.31665
> PRESS(mar1.lm)
[1] 41.52744
> AIC(mar1.lm)
[1] 54.50941
> PRESS(mar2.lm)
[1] 49.91057
> AIC(mar2.lm)
[1] 58.67148
> PRESS(mar3.lm)
[1] 91.23632
> AIC(mar3.lm)
[1] 67.5824
> PRESS(mar4.lm)
```

```
[1] 48.24616
> AIC(mar4.lm)
[1] 58.34142
> PRESS(mar5.lm)
[1] 102.1150
> AIC(mar5.lm)
[1] 68.25571
> PRESS(mar6.lm)
[1] 79.66034
> AIC(mar6.lm)
[1] 65.64033
> PRESS(mar7.lm)
[1] 87.67762
> AIC(mar7.lm)
[1] 67.18876
```

These results are summarized in the table below.

	Pred Alt Lat Whistles	Pred Alt Whistles	Lat Whistles	Alt Lat Whistles	Pred Whistles	Lat Whistles	Alt Whistles	Whistles
AIC	56.31	54.51	58.67	67.58	58.34	68.26	65.64	67.19
PRESS	46.20	41.53	49.91	91.24	48.25	102.12	79.66	87.68

These results are consistent with the expectation that the predator index should be used with whistles. The models without the predator index tend to do worse than those without. Dropping either predator index or whistles makes the other perform poorly. The "best" model here includes predator index, whistles, and altitude, however predator index and whistles alone might be a reasonable choice.

17.7 Sample size dependency

Suppose there are two possible predictors, with 1000 data points. The relation between the data and response is strong for the first predictor, but weaker for the second (the formula used to generate the data is $y = x_1 + 0.05x_2$, making $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 = 0.05$. Using AIC on the full dataset does show both x_1 and x_2 should be kept as predictors.

```
> set.seed(1)
> x1<-rnorm(1000,0,10)
> x2<-rnorm(1000,0,10)
> y<-rnorm(1000,x1+.1*x2,8)
>
> ## Use AIC to find the best model
> sim.lm1<-lm(y~x1+x2)
> step(sim.lm1)
```

```
Start:  AIC=4222.21
y ~ x1 + x2

       Df Sum of Sq    RSS    AIC
<none>               67776   4222
- x2    1      1495  69271   4242
- x1    1    115555 183331   5215

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
     0.1299       1.0392       0.1176
```

Now suppose only the first 30 data were observed. Using the same process as before, AIC now indicates that x_2 should be dropped.

```
> sim.lm2<-lm(y[1:30]~x1[1:30]+x2[1:30])
> step(sim.lm2)
Start:  AIC=134.79
y[1:30] ~ x1[1:30] + x2[1:30]

           Df Sum of Sq    RSS    AIC
- x2[1:30]  1     146.1 2341.4  134.7
<none>                  2195.3  134.8
- x1[1:30]  1    2397.4 4592.7  154.9

Step:  AIC=134.72
y[1:30] ~ x1[1:30]

           Df Sum of Sq    RSS    AIC
<none>                  2341.4  134.7
- x1[1:30]  1    2505.6 4846.9  154.5

Call:
lm(formula = y[1:30] ~ x1[1:30])

Coefficients:
(Intercept)     x1[1:30]
     -1.831        1.006
```

This same pattern holds true for PRESS statistics as well. Note, however, that the inclusion of x_2 does not mean that x_2 is significant.

```
> summary(sim.lm2)
```

```
Call:
lm(formula = y[1:30] ~ x1[1:30] + x2[1:30])

Residuals:
     Min       1Q   Median       3Q      Max
-15.2563  -5.9555  -0.5878   6.2630  20.1907

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3838     1.6863  -0.821    0.419
x1[1:30]      0.9869     0.1817   5.430 9.61e-06 ***
x2[1:30]      0.2157     0.1609   1.341    0.191
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.017 on 27 degrees of freedom
Multiple R-Squared: 0.5471,     Adjusted R-squared: 0.5135
F-statistic: 16.31 on 2 and 27 DF,  p-value: 2.271e-05
```

Lecture 18

Extensions

There are many ways to extend the material in this course to other applications and more general cases. The purpose of this lecture is not to give a comprehensive look at these topics, but rather a cursory look; In Loveday's words, "enough to get into trouble." The scope of statistics in ecology is quite extensive, so even these necessarily leave out some topics.

18.1 General linear models

One obvious extension of the models discussed so far are to have a multivariate responses. Often the response is multivariate. The example in section 5.8.2, for instance, is really a multivariate response. The response in that example was taken to be percent iron, however the data which were actually collected were several measurements of different metals. Combining all of these gives a more powerful test to show significant differences between the sites.

The setup for general linear models (or the special case of categorical predictors only, manova) should look familiar. Instead of a vector y for the response, we have several response vectors, y_1, y_2, etc. These are commonly combined into a response matrix Y. In the archaeological example above, for instance, each column of the matrix is the measurements of percent of a particular metal.

$$Y = \left(\begin{array}{c|c|c} y_1 & y_2 & y_3 \end{array} \right) \tag{18.1}$$

To make the dimensions match, then, the parameters β must also be a matrix. The model, in matrix form, for general linear models looks very similar to the univariate case. The errors are likewise a random matrix of appropriate dimension, however unlike ols, these can be correlated in certain ways.

$$Y = X\beta + \epsilon \tag{18.2}$$

Solving for β using maximum likelihood (or least squares) as before, yields $\beta = (X'X)^{-1}X'Y$.

One of the very common methods of gaining inference is simple likelihood ratio tests. This, for general linear models, is termed "Wilk's lambda statistic." There are several other methods of testing the model, as well, however.

18.1.1 Archaeological metals

The full dataset from the archaeological sites in section 5.8.2 is shown in the table below. Recall that using Iron alone, it was evident that there were two different sites.

Al	Fe	Mg	Ca	Na	Site
14.4	7.00	4.30	0.15	0.51	L
13.8	7.08	3.43	0.12	0.17	L
14.6	7.09	3.88	0.13	0.20	L
11.5	6.37	5.64	0.16	0.14	L
13.8	7.06	5.34	0.20	0.20	L
10.9	6.26	3.47	0.17	0.22	L
10.1	4.26	4.26	0.20	0.18	L
11.6	5.78	5.91	0.18	0.16	L
11.1	5.49	4.52	0.29	0.30	L
13.4	6.92	7.23	0.28	0.20	L
12.4	6.13	5.69	0.22	0.54	L
13.1	6.64	5.51	0.31	0.24	L
12.7	6.69	4.45	0.20	0.22	L
12.5	6.44	3.94	0.22	0.23	L
11.8	5.44	3.94	0.30	0.04	C
11.6	5.39	3.77	0.29	0.06	C
18.3	1.28	0.67	0.03	0.03	I
15.8	2.39	0.63	0.01	0.04	I
18.0	1.50	0.67	0.01	0.06	I
18.0	1.88	0.68	0.01	0.04	I
20.8	1.51	0.72	0.07	0.10	I
17.7	1.12	0.56	0.06	0.06	A
18.3	1.14	0.67	0.06	0.05	A
16.7	0.92	0.53	0.01	0.05	A
14.8	2.74	0.67	0.03	0.05	A
19.1	1.64	0.60	0.10	0.03	A

The testing procedure is fairly straight-forward. Use the `cbind()` command to group the response variables.

```
> met<-read.table(
+     file="http://students.washington.edu/nesse/qerm514/data/metals.txt",
+     header=T)
> met.lm<-lm(cbind(Fe,Mg,Al,Ca,Na)~Site,data=met)
> anova(met.lm,test="Wilk")
Analysis of Variance Table

               Df  Wilks approx F num Df den Df    Pr(>F)
(Intercept)  1.000   0.01   523.07  5.000 18.000 < 2.2e-16 ***
Site         3.000   0.01    13.09 15.000 50.091 1.840e-12 ***
Residuals   22.000
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> summary(met.lm)
Response Fe :

Call:
lm(formula = Fe ~ Site, data = met)

Residuals:
     Min       1Q   Median       3Q      Max
-2.11214 -0.33954  0.01143  0.49036  1.22800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5120     0.3155   4.792 8.73e-05 ***
SiteC         3.9030     0.5903   6.612 1.20e-06 ***
SiteI         0.2000     0.4462   0.448    0.658
SiteL         4.8601     0.3676  13.222 6.04e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7055 on 22 degrees of freedom
Multiple R-Squared: 0.9246,     Adjusted R-squared: 0.9143
F-statistic: 89.88 on 3 and 22 DF,  p-value: 1.679e-12


Response Mg :

Call:
lm(formula = Mg ~ Site, data = met)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39643 -0.35893 -0.00500  0.07975  2.40357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6060     0.3745   1.618 0.119901
SiteC         3.2490     0.7007   4.637 0.000127 ***
SiteI         0.0680     0.5297   0.128 0.899011
SiteL         4.2204     0.4363   9.673 2.20e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8375 on 22 degrees of freedom
Multiple R-Squared: 0.8701,     Adjusted R-squared: 0.8524
F-statistic: 49.12 on 3 and 22 DF,  p-value: 6.452e-10
```

```
Response Al :

Call:
lm(formula = Al ~ Site, data = met)

Residuals:
     Min       1Q   Median       3Q      Max
-2.52000 -0.87821  0.01786  0.94393  2.62000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3200     0.6626  26.141  < 2e-16 ***
SiteC        -5.6200     1.2395  -4.534 0.000164 ***
SiteI         0.8600     0.9370   0.918 0.368664
SiteL        -4.7557     0.7719  -6.161 3.35e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.482 on 22 degrees of freedom
Multiple R-Squared: 0.7843,     Adjusted R-squared: 0.7549
F-statistic: 26.67 on 3 and 22 DF,  p-value: 1.627e-07


Response Ca :

Call:
lm(formula = Ca ~ Site, data = met)

Residuals:
      Min        1Q    Median        3Q       Max
-0.082143 -0.022107 -0.002143  0.015393  0.107857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05200    0.02163   2.404   0.0251 *
SiteC        0.24300    0.04047   6.004 4.83e-06 ***
SiteI       -0.02600    0.03060  -0.850   0.4046
SiteL        0.15014    0.02520   5.957 5.38e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.04838 on 22 degrees of freedom
Multiple R-Squared: 0.799,      Adjusted R-squared: 0.7716
F-statistic: 29.16 on 3 and 22 DF,  p-value: 7.546e-08
```

```
Response Na :

Call:
lm(formula = Na ~ Site, data = met)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11071 -0.04571 -0.01400  0.00500  0.28929

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04800    0.04256   1.128 0.271598
SiteC        0.00200    0.07963   0.025 0.980189
SiteI        0.00600    0.06020   0.100 0.921505
SiteL        0.20271    0.04959   4.088 0.000487 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.09518 on 22 degrees of freedom
Multiple R-Squared: 0.5644,     Adjusted R-squared: 0.505
F-statistic: 9.503 on 3 and 22 DF,  p-value: 0.0003209
```

Note that there are 20 parameters being fit in this model. The anova table does indicate, using Wilk's test, that there are some significant differences between sites. It is possible to set up some more sophisticated tests to check if the sites show the same pattern as the iron-only response model (sites A and I being similar and C and L being similar). But the general appearance of this pattern is evident in the significance of the parameters. Note that the base level is site A, so the lack of significance of all the Site I predictors indicates the same pattern generally holds true.

18.2 Multinomial glm

The natural extension of the binomial glm is the multinomial glm. Here there is a multivariate response to the predictors, and there are multiple probabilities p_{ij} (recall that each portion of the response must be assigned a probability such that all the possible responses sum to 1). Just as in the general linear model case, there are separate coefficients β for each of the probabilities. Define $X_i\beta_j = \eta_{ij}$, and write the each p_{ij} as the analogue of the inverse logit transform.

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \sum_{t \neq j} \exp(\eta_{it})} \tag{18.3}$$

This sort of model can be used in cases where there are many possible responses to a question, not just two as was the case in the binomial. Think back to the problem on Brussles sprouts, section 14.5.1. There the data were grouped into "repulsive" and "delicious." The original survey, however, had five possible responses ranging from "very repulsive" to "especially delicious." These data could be modeled using a multinomial

18.3 Generalized Linear Mixed Models

Linear mixed effects models incorporated random effects (wherein some of the parameters were themselves random variables). This class has also looked at generalized linear models, where the responses are non-normal. The combination of these two ideas gives rise to generalized linear mixed models, or glmm. These arise from the incorporation of random effects in generalized linear models.

An example might include a randomized block design experiment where the response variable is a count or proportion modeled with a poisson or binomial. The random effect is still modeled to be normal (although some literature exists on estimating non-normal random effects as well).

Estimating the parameters of glmms can be done in several ways. According to Faraway, there is no consensus on what the optimal method is for estimating these models. It is still an active area of research.

18.4 Ridge regression

One topic not carefully dealt with in this course is the common problem of correlated predictors. Correlated predictors give rise to highly unstable estimates of $\hat{\beta}$, see example 4.7.2. One way of thinking about the instability is to consider the ols solution, $\hat{\beta} = (X'X)^{-1}X'y$. If two predictors are highly correlated, then $X'X$ is nearly singular (determinant is nearly zero–if the predictors are perfectly correlated, then the $X'X$ *is* singular and thus has no inverse). The inverse of a nearly singular matrix is likely to be highly sensitive to the values of X. This problem can be reduced, therefore, with a modification to the $X'X$ matrix, to make it (in a sense) less almost singular. That is, move the determinant away from zero. This will also have a nice geometric interpretation as well.

To modify $X'X$ to make it less singular (after normalizing by subtracting the mean of each column from the columns of X), how about using $X'X + \lambda I$. This should increase the size of the determinants (and eigenvalues) away from zero. Thus the ridge $\hat{\beta}$ can be found using

$$\hat{\beta}_{\text{Ridge}} = (X'X + \lambda I)^{-1}X'y \tag{18.4}$$

These solutions can be thought of as restricting the size of $\hat{\beta}'\hat{\beta}$. Ridge regression finds the best estimate of β subject to the condition that β is not too large (as measured by $\hat{\beta}'\hat{\beta}$).

Generally the values of λ are fairly small. There are automated methods for picking λ, together with ridge regression algorithms implemented in the `MASS` package using the commands `lm.ridge()` and `select()` for the selection of λ.

18.5 Lasso regression

A method related to ridge regression is ridge regression is lasso regression. Rather than looking at keeping $\beta'\beta$ below some constant, lasso regression fits the linear model under the assumption that $\sum_{i=1}^{p} |\beta_i| < t$ for some constant t (again, after normalizing X). Note that the constant term β_0 is not subject to the constraint.

Lasso regression works in a similar manner to ridge regression, however the coefficients are often exactly zero.

18.6 Generalized additive models

In general, statistical models are of the form $y_i = f(x_{i1}, \ldots, x_{ik}, Z)$, where Z is a random variable. There are too many possible functions f to determine the optimal model from the data alone.[1] The general method of modeling has been to restrict the possible functions to a subset of all possible models.

Generalized additive models considers models of the form

$$y_i = g_1(x_{i1}) + g_2(x_{i2}) + \cdots + g_k(x_{ik}) + \epsilon_i \tag{18.5}$$

Each of the functions g is unknown, but assumed to be continuous and smooth, and has to be estimated numerically. In the one-predictor case, this is simply fitting a smoothed line which goes thorough the data (commonly just a moving average). Multivariate regression of this form is numerically more difficult, but can also be carried out.

18.7 Nonlinear mixed effects models

A natural way to model nonlinear responses to predictors is to assume a form of the model and consider the parameters to be random variables. Contrast this against the nonlinear least squares approach, which assumes that errors are observational.

These sorts of models are very natural in a wide variety of ecological situations.[2] Just as in the orthodontic data, where an individual's growth rate was modeled as a random effect, it may make sense to model growth as a random effect in a nonlinear growth model.

An example, fitting a von Bertalanffy curve to fish growth data, as has been done many times in class, attempts to fit three parameters: L_∞, k, and t_0. One approach to modeling these data (used, for example, in section 2.8.5) is a nonlinear least squares model. This assumes each fish is growing on an identical curve, and all the variability of the data comes in from errors in measuring the length.[3] A more natural guess at the source of the variation is to assume each fish has a slightly different growth curve, the parameters of which are drawn from a random distribution.

The traditional nonlinear mixed effects model assumes the parameters each have a normal distribution (although they may be correlated). The ideal circumstance, like for the linear mixed effects models, is to have repeated measures of the same fish (something which rarely occurs in real fisheries settings other than, perhaps, mark recapture experiments).

18.7.1 Ageing fish

The mysterious and elusive estimator fish (*E. statisticsii*, see figure **??**) has been the subject of an extensive measurement to determine its growth parameters for the von Bertalanffy model.

Each of 100 fish has been measured three times. Since the goal here is to estimate parameters, the actual parameters are noted below. The three parameters k, L_∞, and t_0 are all normally distributed. The parameters are given in the table below. Additionally, an observational error, normal with standard deviation of 3, is added to the length measurement (and all lengths are set to be positive).

[1] This is related to a problem in the philosophy of science, under the name "Duhem-Quine underdetermination."

[2] Arguably, more applicable than their use would indicate, although they are commonly used.

[3] Interestingly enough, the length at capture is probably one one quantity we are the most sure of. Ageing fish (determining the age) is generally a more uncertain process than measuring length.

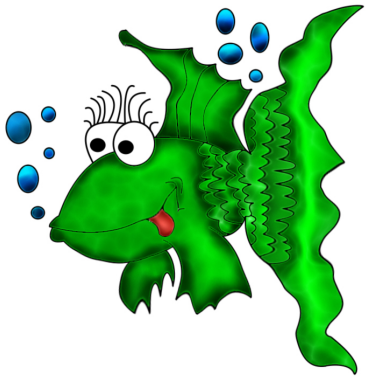


Figure 18.1: A fish, from User:ArielGold, http://commons.wikimedia.org/wiki/Image:Cartoon-Fish.jpg. GFDL 1.2

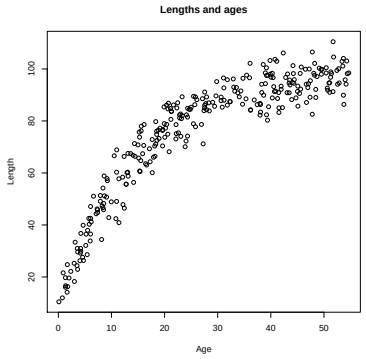
	Mean	Stan. Dev.
k	0.07	0.001
L_∞	100	3
t_0	1	0.1

The commands to create the data are shown below.

```
> library(nlme)
> vonbert<-function(age,k,LInf,t0){
    return(LInf*(1-exp(-k*(age+t0))))
  }
> set.seed(1)
> ages<-round(runif(100,0,55),2)
> len1<-NULL
> len2<-NULL
> len3<-NULL
> age1<-NULL
> age2<-NULL
> age3<-NULL
> for(tt in 1:length(ages)){
      k<-rnorm(1,.07,.01)
      LInf<-rnorm(1,100,3)
      t0<-rnorm(1,1,.1)
      age.tmp<-round(runif(1,0,55),2)
      len1[tt]<-max(0,rnorm(1,vonbert(age.tmp,k,LInf,t0),3))
      age1[tt]<-age.tmp
      age.tmp<-round(runif(1,0,55),2)
      len2[tt]<-max(0,rnorm(1,vonbert(age.tmp,k,LInf,t0),3))
      age2[tt]<-age.tmp
      age.tmp<-round(runif(1,0,55),2)
```

```
      len3[tt]<-max(0,rnorm(1,vonbert(age.tmp,k,LInf,t0),3))
      age3[tt]<-age.tmp
  }
>
> subj<-as.factor(1:length(ages))
> len.df<-data.frame(c(age1,age2,age3),c(len1,len2,len3),rep(subj,3))
> names(len.df)<-c("ages","len","subj")
> len.gd<-groupedData(len~ages|subj,data=len.df)
> plot(len.df$ages,len.df$len,
      main="Lengths and ages",
      xlab="Age",
      ylab="Length")
```

The data are really good, as far as length-age data go. There are three hundred observations.



We could ignore that there are repeated observations on the same individual and fit a nonlinear least squares model. Alternatively, taking advantage of the grouping, we could fit a nlme model.

NLME model

The syntax for nlme models is obscure and hard to follow.[4]

```
> len.nlme<-nlme(len~vonbert(ages,a,b,c),
      fixed=a+b+c~1,
      random=a+b+c~1,
      start=c(a=.05,b=85,c=1.5),
      data=len.gd)
> summary(len.nlme)
Nonlinear mixed-effects model fit by maximum likelihood
  Model: len ~ vonbert(ages, a, b, c)
 Data: len.gd
       AIC     BIC    logLik
  608.2232 645.261 -294.1116
```

[4]Perhaps this is my lack of experience on these sorts of models, but it took me an hour to figure out what worked, mostly by trial and error.

```
Random effects:
 Formula: list(a ~ 1, b ~ 1, c ~ 1)
 Level: subj
 Structure: General positive-definite, Log-Cholesky parametrization
         StdDev       Corr
a        9.989536e-03 a      b
b        3.240630e+00  0.037
c        1.041150e-01 -0.018 -0.089
Residual 1.793420e-05

Fixed effects: a + b + c ~ 1
      Value Std.Error  DF   t-value p-value
a   0.07029 0.0010040 198  70.01428       0
b 100.18668 0.3256956 198 307.60831       0
c   0.99842 0.0104762 198  95.30376       0
 Correlation:
  a      b
b  0.037
c -0.019 -0.089

Standardized Within-Group Residuals:
          Min             Q1           Med             Q3           Max
-1.665666e-01 -1.306982e-04 -3.580606e-07  1.504902e-04  8.502513e-02

Number of Observations: 300
Number of Groups: 100
```

The estimation of the fixed effects worked pretty well. The standard values for k, L_∞ and t_0 are 0.07029, 100.18668, and 0.99842 respectively. The estimated between-fish variation in these parameters (measured in standard deviation) is 0.0009989, 3.24, and 0.0104. All of these values are close to the true values used to generate the data.

NLS

If the group structure is ignored, it is also possible to estimate these data using the nonlinear least squares method.

```
> len.nls<-nls(len~vonbert(ages,a,b,c),
+     start=list(a=.05,b=85,c=1.5),
+     data=len.df)
> summary(len.nls)

Formula: len ~ vonbert(ages, a, b, c)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
```

```
a  0.068234   0.002651  25.737  < 2e-16 ***
b 99.276479   0.890601 111.471  < 2e-16 ***
c  1.344134   0.275931   4.871 1.81e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.629 on 297 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 1.965e-07
```

The nls gives a pretty good estimate of the parameters, although not as good as the nlme. The standard errors, however, are not immediately interpretable. Partially they reflect the variability of the parameters among fish, and partially they reflect the observational error.

Appendix A

Notation

A great deal of notation is used in mathematics and statistics, often it clarifies, sometimes it obfuscates. For reasons which are somewhat mysterious to me, in statistics

1. Theorems, rules, ideas, etc. are almost never named after the people who discovered or invented them.[1]

2. Often, several (or many) wildly different notational conventions (or sometimes deceptively similar) conventions exist for the same idea.

To cut through some of that, this is an appendix which covers some of the notation used.

A.1 Mathmatical notation

A.1.1 Numbers sums and products

To denote the set of all real numbers,[2] the symbol $\mathbb{R}$ is commonly used in these notes, although $\mathbf{R}$ and $\mathfrak{R}$ are sometimes used in other books. The set of all vectors of length n is usually denoted $\mathbb{R}^n$. The collection of all integers (whole numbers, both positive and negative) is $\mathbb{Z}$, while the nonnegative numbers is usually $\mathbb{Z}^+$ or in some places $\mathbb{N}$.[3]

The element symbol, $\in$, is used to indicate when a particular number is in a set (although it is not used for matrices), for instance $a \in \mathbb{R}$ means a is a real number (it is in the set of one dimensional real numbers).

The sum of a collection of numbers, $a_1, a_2, \ldots, a_n$ is usually written

$$a_1 + a_2 + \cdots + a_n = \sum_{i=1}^{n} a_i. \tag{A.1}$$

[1] This is sometimes formulated as *Stigler's law of eponymy*: "No scientific discovery is named after its original discoverer." It is not surprising that Stigler was a statistician.

[2] As opposed to complex numbers or worse.

[3] In many cases, either of these can be used to mean only the positive numbers as well.

Likewise the product of these numbers is usually written

$$a_1 a_2 \cdots a_n = \prod_{i=1}^{n} a_i. \tag{A.2}$$

A.1.2 Matrix notation

For a matrix A, the constituent parts are called "elements" and are denoted with a lower case a_{ij} for the ith row and jth column. Vectors have varying notation; in these notes, a lower case letter is used (which can make it easy to confuse with ordinary numbers). Other books commonly use bold for matrices and vectors, or sometimes $\vec{a}$ for vectors.

Some common notation used for matrix operations (the † denotes an operation not used in this course)

Name	notation	alt. notation	Name	notation	alt. notation
Transpose	A'	A^T	Determinant	$\|A\|$	$\det(A)$
Trace	$\mathrm{tr}(A)$		Rank	$\mathrm{rank}(A)$	
Inverse	A^{-1}		Identity matrix	I_n	I
Kronecker Product†	$A \otimes B$		Hadamard product†	$\odot$	
Moore-Penrose inv.†	A^+				

A.2 Probability/Likelihood

Random variables are usually denoted with a capital letter, such as X. The pdf or pmf for X is denoted $f_X(x)$ (the letter f is most commonly used, but it could be anything lowercase), and when the random variable under consideration is clear from context, the X is dropped, such as $f(x)$. To explicitly indicate the parameters being used in the distribution, a vertical bar is used: $f_X(x|\theta)$ means the distribution of random variable X, which depends on the value of the parameter(s) θ. Most often, in these notes and elsewhere, Greek letters are used to indicate parameters.[4]

Many books will use $l(\theta|x)$ to mean the same thing as $f(x|\theta)$ when speaking as a function of θ. This is done for (silly) philosophical reasons, and is not done in these notes (even though most books do this).

Although they don't come up frequently in this class, the cumulative density function of a random variable X is usually denoted with a capital letter corresponding to the lowercase pdf/pmf. That is $F_X(x)$ is the cdf of X, which has pdf $f_X(x)$. Since the normal distribution is so commonly used, it gets its own notation: The cdf of a standard normal is a capital phi, $\Phi(x)$, while the pdf is a lowercase phi, $\phi(x)$ (this convention is universal as far as I have seen).

Probability of an event is denoted here with a $P(event)$, however it varies in notation from book to book. Most commonly, the above notation, $Pr(event)$ or $Pr\{event\}$ are used.

A.3 Commonly used letters

There is generally no set standards for using particular letters to mean particular things, but some usual trends develop (here and elsewhere). Some ways of modifying This list is not comprehensive,

[4] The notable exception, which is nearly always the same in every book, ϵ is used to mean a normal additive error in linear models.

nor is it always universally followed even in these notes.

ϵ	Normal error, or just error.
g	Commonly used for a function, often for glm link functions.
Γ	The gamma function
L	Likelihood function (or log likelihood function).
λ	Likelihood ratio, Poisson rate parameter, fit parameter in a model
η	Often used to be $X\beta$ in glms
θ	An arbitrary vector of parameters
Σ	Sum, Covariance matrix
w	Predictor
x	Predictor, (capital) Pearson's statistic
y	Response
Z	Normal random variable