Advanced linear models for data science

Brian Caffo

January 6, 2017

Contents

Preface

This book is written as a companion book to the Advanced Linear Models for Data Science Coursera class. Also check out the Data Science Specialization by Brian Caffo, Roger Peng and Jeff Leek. However, if you do not take the class, the book mostly stands on its own. A useful component of the book is a series of [LINK] YouTube videos that comprise the Coursera class.

The book is intended to be a low cost introduction to the important field of advanced linear models. The intended audience are students who are numerically and computationally literate, have taken a course on statistical inference, have taken a regression class, can program in R and have a fairly high level of mathematical sophistication including: linear algebra, multivariate calculus and some proof-based mathematics. The book is offered for free with variable pricing (html, pdf, epub, mobi) on LeanPub.

Chapter 1

Introduction

Linear models are the cornerstone of statistical methodology. Perhaps more than any other tool, advanced students of statistics, biostatistics, machine learning, data science, econometrics, etcetera should spend time learning the finer grain details of this subject.

In this book, we give a brief, but rigorous treatment of advanced linear models. It is advanced in the sense that it is of level that an introductory PhD student in statistics or biostatistics would see. The material in this book is standard knowledge for any PhD in statistics or biostatistics.

1.1 Prerequisites

Students will need a fair amount of mathematical prerequisites before trying to undertake this class. First, is multivariate calculus and linear algebra. Especially linear algebra, since much of the early parts of linear models are direct applications of linear algebra results applied in a statistical context. In addition, some basic proof based mathematics is necessary to follow the proofs.

We will also assume some basic mathematical statistics. The courses Mathematical Biostatistics Boot Camp 1 and Mathematical Biostatistics Boot Camp 2 by the author on Coursera would suffice. The Statistical Inference is a lower level treatment that with some augmentated reading would also suffice. There is a Leanpub book for this course as well.

Some basic regression is necessary. The Regression Models also by the author would suffice. Note that there is a Leanpub book for this class.

Chapter 2

Background

Watch this video before beginning.

Before we begin we need a few matrix prerequisites. Let, $f : \mathbb{R}^p \rightarrow \mathbb{R}$, be a function from the p dimensional real line to the real line. Assume that f is linear, i.e. $f(\mathbf{x}) = \mathbf{a}^t\mathbf{x}$ Then $\nabla f = \mathbf{a}$. That is, the function that contains the elementwise derivatives of f (the gradient) is constant with respect to x.

Consider now the matrix $\mathbf{A}$ ($p \times p$) and the quadratic form:

$$f(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x}.$$

Then $\nabla f = 2\mathbf{A}^t\mathbf{x}$. The second derivative matrix (Hessian) (where the i, j element is the derivative with respect to the i and j elements of this vector) is then $2\mathbf{A}$.

2.1 Example

Consider an example that we will become very familiar with, **least squares**. Let $\mathbf{y}$ be an array of dimension $n \times 1$ and $\mathbf{X}$ be an $n \times p$ full rank matrix. Let $\boldsymbol{\beta}$ be a p vector of unknowns. Consider trying to minimize the function:

$$f(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^t\mathbf{y} - 2\mathbf{y}^t\mathbf{X}^t\boldsymbol{\beta} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}.$$

The gradient of f (with respect to $\boldsymbol{\beta}$) is:

$$\nabla f(\boldsymbol{\beta}) = -2\mathbf{X}\mathbf{y} + \mathbf{X}^t\mathbf{X}\boldsymbol{\beta}. \tag{2.1}$$

A useful result is that if $\mathbf{X}$ is of full column rank then $\mathbf{X}^t\mathbf{X}$ is square and full rank and hence invertible. Thus, we can calculate the root of the gradient (2.1) and obtain the solution:

$$\boldsymbol{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}.$$

We will talk about these solutions at length. Also, it remains necessary to show that this is a minimum and not just an inflection point. We can do this by checking a second derivative condition. Taking the second derivative of (2.1) we get

$$\mathbf{X}^t\mathbf{X},$$

a positive definite matrix. Thus our solution is indeed a minimum.

2.1.1 Coding example

Watch this video before beginning.

Load the `mtcars` dataset in R with `data(mtcars)`. Let's set our $\mathbf{y}$ vector as `y = mtcars$mpg` and our $\mathbf{X}$ matrix as a vector of ones, horsepower and weight: `x = cbind(1, mtcars$hp, mtcars$wt)`. Let's find the value of $\boldsymbol{\beta}$ that minimizes $||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$.

```
> y = mtcars$mpg
> x = cbind(1, mtcars$hp, mtcars$wt)
> solve(t(x) %*% x) %*% t(x) %*% y
            [,1]
[1,] 37.22727012
[2,] -0.03177295
[3,] -3.87783074
> # Compare with the estimate obtained via lm
> coef(lm(mpg ˜ hp + wt, data = mtcars))
(Intercept)          hp          wt
37.22727012 -0.03177295 -3.87783074
```

2.2 Averages

Watch this video before beginning.

Consider some useful notational conventions we'll use. $\mathbf{1}_n$ is an n vector containing only ones while $\mathbf{1}_{n\times p}$ is an $n \times p$ matrix containing ones. $\mathbf{I}$ is the identity matrix, which we will subscript if a reminder of the dimension is necessary.

We denote by $\bar{y}$ the average of the n vector $\mathbf{y}$. Verify for yourself that $\bar{y} = \frac{1}{n}\mathbf{y}^t\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n^t\mathbf{y}$. Furthermore, note that $(\mathbf{1}_n^t\mathbf{1}_n)^{-1} = \frac{1}{n}$ so that

$$\bar{y} = (\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\mathbf{y}.$$

Consider our previous least squares problem. If $\mathbf{X} = \mathbf{1}_n$ and $\boldsymbol{\beta}$ is just the scalar β. The least squares function we'd like to minimize is:

$$(\mathbf{y} - \mathbf{1}_n\beta)^t(\mathbf{y} - \mathbf{1}_n\beta) = ||\mathbf{y} - \mathbf{1}_n\beta||^2.$$

Or, what constant vector best approximates $\mathbf{y}$ it the terms of minimizing the squared Euclidean distance? From above, we know that the solution is of the form

$$\boldsymbol{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

which in this specific case works out to be:

$$\bar{y} = (\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\mathbf{y}.$$

That is, the average is the best scalar estimate to minimize the Euclidean distance.

2.3 Centering

Continuing to work with our vector of ones, note that

$$\mathbf{y} - \mathbf{1}_n \bar{y}$$

is the centered version of $\mathbf{y}$ (in that it has the mean subtracted from each element of the $\mathbf{y}$ vector). We can rewrite this as:

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{1}_n\bar{y} = \mathbf{y} - (\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}^t\mathbf{1}\mathbf{y} = \left\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}^t\right\}\mathbf{y}.$$

In other words, multiplication by the matrix $\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}^t\}$ centers vectors. To check that $\tilde{\mathbf{y}}$ is centered, consider multiplying by $\mathbf{1}_n$ (which sums its elements).

$$\mathbf{1}_n^t\tilde{\mathbf{y}} = \mathbf{1}_n^t\left\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\right\}\mathbf{y} = \left\{\mathbf{1}_n^t - \mathbf{1}_n^t\mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\right\} = \left\{\mathbf{1}_n^t - \mathbf{1}_n^t\right\}\mathbf{y} = 0.$$

This operation can be very handy for centering matrices. For example, if $\mathbf{X}$ is an $n \times p$ matrix then the matrix $\tilde{\mathbf{X}} = \{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}^t\}\mathbf{X}$ is the matrix with every column centered. Conversely, right multiplication by $\mathbf{I} - \mathbf{1}_p(\mathbf{1}_p^t\mathbf{1}_p)^{-1}\mathbf{1}_p^t$ centers every row of $\mathbf{X}$.

2.3.1 Coding example

Watch this video before beginning.

Let's take our $\mathbf{X}$ matrix defined previously from the `mtcars` dataset and mean center it. We'll contrast using matrix manipulations versus (preferable) R functions.

```
> n = nrow(x)
> I = diag(rep(1, n))
> H = matrix(1, n, n) / n
> xt = (I - H) %*% x
> apply(xt, 2, mean)
[1] 0.000000e+00 0.000000e+00 2.168404e-16
> ## Doing it using sweep
> xt2 = sweep(x, 2, apply(x, 2, mean))
> apply(xt2, 2, mean)
[1] 0.000000e+00 0.000000e+00 3.469447e-17
```

2.4 Variance

Watch this video before beginning.

The standard sample variance is the average deviation of the observations from the sample mean (usually using $n-1$ rather than n). That is,

$$S^2 = \frac{1}{n-1}||\mathbf{y} - \mathbf{1}_n\bar{y}||^2 = \frac{1}{n-1}\tilde{\mathbf{y}}^t\tilde{\mathbf{y}}$$

We can write out the norm component of this as:

$$\mathbf{y}^t \left\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\right\} \left\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\right\} \mathbf{y} = \mathbf{y}^t \left\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\right\} \mathbf{y}.$$

Thus, our sample variance is a fairly simple quadratic form. Notice the fact that the matrix $\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\}$ is both symmetric and idempotent.

Similarly, if we have two vectors, $\mathbf{y}$ and $\mathbf{z}$ then $\frac{1}{n-1}\mathbf{y}\left\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\right\}\mathbf{z}$ is the empirical covariance between them. This is then useful for matrices. Consider that

$$\frac{1}{n-1}\mathbf{X}^t \left\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\right\} \mathbf{X} = \frac{1}{n-1}\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$$

is a matrix where each element is the empirical covariance between columns of $\mathbf{X}$. This is called the variance/covariance matrix.

2.4.1 Coding example

Watch this video before beginning.

Let's manually calculate the covariance of the `x` matrix from before.

```
> n = nrow(x)
> I = diag(rep(1, n))
> H = matrix(1, n, n) / n
> round(t(x) %*% (I - H) %*% x / (n - 1), 6)
     [,1]       [,2]      [,3]
[1,]    0    0.00000  0.000000
[2,]    0 4700.86694 44.192661
[3,]    0   44.19266  0.957379
> var(x)
     [,1]       [,2]      [,3]
[1,]    0    0.00000  0.000000
[2,]    0 4700.86694 44.192661
[3,]    0   44.19266  0.957379
```

Recall, the first column was all ones; thus the row and column of zeros in the variance.

Chapter 3

Single parameter regression

3.1 Mean only regression

Consider least squares where we only want horizontal lines. Let our outcome be $\mathbf{y} = (y_1, \ldots, y_n)^t$ and recall that $\mathbf{1}_n$ is an n vector of ones. We want to minimize $f(\mu) = ||\mathbf{y} - \mathbf{1}\mu||^2$ with respect to μ.

Taking derivatives with respect to μ we obtain that

$$\frac{df}{d\mu} = -2n\bar{y} + 2n\mu.$$

This has a root at $\hat{\mu} = \bar{y}$. Note that the second derivative is $2n > 0$. Thus, the average is the least squares estimate in the sense of minimizing the Euclidean distance between the observed data and a constant vector. We can think of this as projecting our n dimensional onto the best 1 dimensional subspace spanned by the vector $\mathbf{1}$. We'll rely on this form of thinking a lot throughout the text.

3.2 Coding example

Let's use the `diamond` dataset

```
> library(UsingR); data(diamond)
> y = diamond$price; x = diamond$carat
> mean(y)
[1] 500.0833
> #using least squares
> coef(lm(y ~ 1))
[1] 500.0833
```

Thus, in this example the mean only least squares estimate obtained via `lm` is the empirical mean.

3.3 Regression through the origin

Watch this video before beginning.

Let $\mathbf{x} = (x_1, \ldots, x_n)'$ be another vector. Consider now the regression through the origin problem. We want to minimize $f(\beta) = ||\mathbf{y} - \mathbf{x}\beta||^2$ with respect to β. This is called regression through the origin for the following reason. First note that the pairs, (x_i, y_i), form a scatterplot. Least squares is then finding the best multiple of the $\mathbf{x}$ vector to approximate $\mathbf{y}$. That is, finding the best line of the form $y = \beta x$ to fit the scatter plot. Thus we are considering lines through the origin hence the name regression through the origin.

Notice that $f(\beta) = \mathbf{y}^t\mathbf{y} - 2\mathbf{y}^t\mathbf{x} + \mathbf{x}^t\mathbf{x}$. Then

$$\frac{df}{d\beta} = -2\mathbf{y}'\mathbf{x} + 2\mathbf{x}^t\mathbf{x}$$

Setting this equal to zero we obtain the famous equation:

$$\hat{\beta} = \frac{\mathbf{y}^t\mathbf{x}}{\mathbf{x}^t\mathbf{x}} = \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}$$

We'll leave it up to the reader to check the second derivative condition. Also, we'll leave it up to you to show that the mean only regression is a special case that agrees with the result.

Notice that we have shown the function

$$g : \mathbb{R}^n \to \mathbb{R}$$

defined by $g(\mathbf{y}) = \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}\mathbf{x}$ projects any n dimensional vector $\mathbf{y}$ into the linear space spanned by the single vector $\mathbf{x}$, $\{\beta\mathbf{x} \mid \beta \in \mathbb{R}\}$.

3.4 Centering first

Watch this video before beginning.

A line through the origin is often not useful. Consider centering the $\mathbf{y}$ and $\mathbf{x}$ first. The the origin would be at the mean of the $\mathbf{y}$ vector and the mean of the $\mathbf{x}$ vector. Let $\tilde{\mathbf{y}} = \{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\}\,\mathbf{y}$ and $\tilde{\mathbf{x}} = \{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\}\,\mathbf{x}$. Then regression through the origin (minimizing $||\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\gamma||^2$ for γ) for the centered data yields the solution $\hat{\gamma} = \frac{\langle \tilde{\mathbf{y}}, \tilde{\mathbf{x}} \rangle}{\langle \tilde{\mathbf{x}}, \tilde{\mathbf{x}} \rangle}$. However, from the previous chapter, we know that

$$\langle \tilde{\mathbf{y}}, \tilde{\mathbf{x}} \rangle = \mathbf{y}^t \left\{\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t\right\} \mathbf{x}^t = (n-1)\hat{\rho}_{xy}\sigma_x\sigma_y$$

and similarly $\langle \tilde{\mathbf{x}}, \tilde{\mathbf{x}} \rangle = (n-1)\hat{\sigma}_y^2$. Here, $\hat{\rho}_{xy}$ and $\hat{\sigma}_y^2$ are the empirical correlation and variance, respectively. Thus our regression through the origin estimate is

$$\hat{\gamma} = \rho_{xy}\frac{\sigma_y}{\sigma_x}.$$

That is, the best fitting line that has to go through the center of the data has a slope equal to the correlation times the ratio of the standard deviations. If we reverse the role of $\mathbf{x}$ and $\mathbf{y}$, we simply invert the ratio of the standard deviations. Thus we also note, that if we center and scale our data first so that the resulting vectors have mean 0 and variance 1, our slope is exactly the correlation between the vectors.

3.4.1 Coding example

Watch this video before beginning.

Let's continue with the diamond example. We'll center the variables first.

```
> yc = y - mean(y);
> xc = x - mean(x)
> sum(yc * xc) / sum(xc * xc)
[1] 3721.025
> coef(lm(yc ~ xc - 1))
      xc
3721.025
> cor(x, y) * sd(y) / sd(x)
[1] 3721.025
```

3.5 Bonus videos

Watch these videos before moving on. (I had created them beofre I reorganized chapters.)

Sneak preview of projection logic.
Coding example.
Sneak preview of linear regression.
Sneak preview of regression generalizations.

Chapter 4

Linear regression

4.1 Introduction to linear regression

Watch this video before beginning.

Now let's consider upping the ante to two parameters. Consider the minimizing

$$||\mathbf{y} - (\beta_0 \mathbf{1}_n + \beta_1 \mathbf{x})||^2 \tag{4.1}$$

over β_1 and β_2. Let's think about this in two ways. First, the space

$$\Gamma = \{\beta_0 \mathbf{1}_n + \beta_1 \mathbf{x} \mid \beta_0, \beta_1 \in \mathbb{R}\}$$

is a two dimensional subspace of $\mathbb{R}^n$. Therefore, the least squares equation finds the projection of the observed data point onto two dimensional subspace spanned by the two vectors $\mathbf{1}_n$ and $\mathbf{x}$.

The second way to think about the fit is to consider the scatterplot of points (x_i, y_i). The goal is to find the best fitting line of the form $y = \beta_0 + \beta_1 x$ by minimizing the sum of the squared vertical distances between the points and the fitted line.

Given what we've done already, it's surprisingly easy to minimize (4.1). Consider fixing β_1 and minimizing with respect to β_0.

$$||\mathbf{y} - \beta_1 \mathbf{x} - \beta_0 \mathbf{1}_n||^2$$

Let $\hat\beta_0(\beta_1)$ be the least squares minimum for β_0 for a given β_1. By our results from mean only regression we know that

$$\hat\beta_0(\beta_1) = \frac{1}{n}(\mathbf{y} - \beta_1 \mathbf{x})\mathbf{1}_n = \bar y - \beta_1 \bar x.$$

Therefore, plugging this into the least squares equation, we know that

$$(4.1) \geq ||\mathbf{y} - \bar y \mathbf{1}_n + \beta_1(\mathbf{x} - \bar x \mathbf{1}_n)||^2 = ||\tilde{\mathbf{y}} - \beta_1 \tilde{\mathbf{x}}||^2, \tag{4.2}$$

where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ are the centered versions of $\mathbf{y}$ and $\mathbf{x}$, respectively. We know from the last chapter (4.2) is minimized by

$$\hat\beta_1 = \hat\rho_{xy} \frac{\hat\sigma_y}{\hat\sigma_x}.$$

Plugging this into $\hat{\beta}_0(\hat{\beta}_1)$ we get that

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}.$$

Therefore, the slope estimate from including an intercept is identical to that of regression through the origin after centering the data. The intercept simply forces the line through the average of the Y's and X's.

4.1.1 Coding example

Watch this video before beginning.

```
> library(UsingR)
> data(diamond)
> x = diamond$carat
> y = diamond$price
> beta1 = cor(x, y) * sd(y) / sd(x)
> beta0 = mean(y) - beta1 * mean(x)
> c(beta0, beta1)
[1] -259.6259 3721.0249
> # versus estimate with lm
> coef(lm(y ~ x))
(Intercept)           x
  -259.6259   3721.0249
 > #Centered regression through the origin
 > sum(yc * xc) / sum(xc^2)
[1] 3721.025
```

4.2 Fitted values

Watch this video before beginning.

We define $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^t$ to be the vector of fitted values. Whereas $\mathbf{y}$ lives in $\mathbb{R}^n$, $\hat{\mathbf{y}}$ lives in Γ, the two dimensional linear subspace of $\mathbb{R}^n$ spanned by the two vectors, $\mathbf{1}_n$ and $\mathbf{x}$. We define $\hat{\mathbf{y}}$ as $\hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$. We can think of our least squares as minmizing

$$||\mathbf{y} - \hat{\mathbf{y}}||$$

over all $\hat{\mathbf{y}} \in \Gamma$. The fitted values are the orthogonal projection of the observed data onto this linear subspace.

4.2.1 Coding example

Watch this video before beginning.

Getting the predicted value for $x = 0.20$ (refer to the previous section `diamond` example).

```
> beta0 + beta1 * .20
[1] 484.5791
> predict(lm(y ~ x), newdata = data.frame(x = .2))
       1
484.5791
```

4.3 Residuals

Watch this video before beginning.

Define $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ to be the vector of residuals. Each residual is the vertical distance between $\mathbf{y}$ and the fitted regression line. Thinking geometrically, the residuals are the orthogonal vector pointing to $\mathbf{y}$ from $\hat{\mathbf{y}}$. Least squares can be thought of as minimizing the sum of the squared residuals. The quantity $||\mathbf{e}||^2$ is called the sum of the squared errors while $\frac{1}{n-2}||\mathbf{e}||^2$ is called the mean squared error or the residual variance.

Watch this video of this coding exercise.

```
> yhat = beta0 + beta1 * x
> e = y - yhat
> max(abs(e - resid(lm(y ~ x))))
```

4.4 Extension to other spaces

Watch this video before beginning.

It is interesting to note that nothing we've discussed is intrinsic to $\mathbb{R}^n$. Any space with a norm and inner product and absent of extraordinary mathematical pathologies would suffice. Hilbert spaces are perhaps the most directly extendable.

As an example, let's develop linear regression for a space of (Lebesgue) square integrable functions. That is, let y be in the space of functions from $[0, 1] \rightarrow \mathbb{R}$ with finite squared itegral. Define the inner product as $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$. Consider finding the best multple approximation to y from the function x (also in that space).

Thus, we want to minimize:

$$||y - \beta_1 x||^2 = \int_0^1 \{y(t) - \beta_1 x(t)\}^2 dt.$$

You might have guessed that the solution will be $\hat{\beta} = \frac{\langle y,x\rangle}{\langle x,x\rangle} = \frac{\langle y,x\rangle}{||x||^2}$. Let's show it (knowing that this is the solution):

$$\begin{aligned}
||y-\beta_1 x||^2 &= ||y-\hat{\beta}_1 x+\hat{\beta}_1 x-\beta_1 x||^2 \\
&= ||y-\hat{\beta}_1 \bar{x}||^2-2\langle y-\hat{\beta}_1 x,\hat{\beta}_1 x-\beta_1 x\rangle+||\hat{\beta}_1 x-\beta_1 x||^2 \\
&\geq ||y-\hat{\beta}_1 \bar{x}||^2-2\langle y-\hat{\beta}_1 x,\hat{\beta}_1 x-\beta_1 x\rangle \\
&= ||y-\hat{\beta}_1 \bar{x}||^2-2\hat{\beta}_1\langle y,x\rangle+2\beta_1\langle y,x\rangle+2\hat{\beta}_1^2||x||^2-2\hat{\beta}_1\beta_1||X||^2 \\
&= ||y-\hat{\beta}_1 \bar{x}||^2-2\frac{\langle y,x\rangle^2}{||x||^2}+2\beta_1\langle y,x\rangle+2\frac{\langle y,x\rangle^2}{||x||^2}-2\beta_1\langle y,x\rangle \\
&= ||y-\hat{\beta}_1 x||^2
\end{aligned}$$

Therefore, $\hat{\beta}_1$ is the least squares estimate.

We can extend this to include an intercept. Let j be a function that is constant at 1. Let $\bar{y}=\int_0^1 y(t)dt$ be the average of y over the domain and define $\bar{x}$ similarly. Then consider minimizing (over β_0 and β_1)

$$||y-\beta_0 j-\beta_1 x||^2$$

First, hold β_1 fixed. By our previous result, we have that the minimizer must satisfy:

$$\beta_0 =< y-\beta_1 x, j > /||j||^2=\bar{y}-\beta_1\bar{x}.$$

Plugging this back into our least squares equation we obtain that:

$$\begin{aligned}
||y-\beta_0 j-\beta_1 x||^2 &\geq ||y-\bar{y}-\beta_1(x-\bar{x})||^2 \\
&= ||\tilde{y}-\beta_1\tilde{x}||^2
\end{aligned}$$

where $\tilde{y}$ and $\tilde{x}$ are the centered functions. We know that this is minimized by

$$\hat{\beta}_1=\frac{\langle\tilde{y},\tilde{x}\rangle}{||\tilde{x}||^2}=\rho_{xy}\frac{\sigma_y}{\sigma_x}.$$

where $\rho_{xy}=\int_0^1\{y(t)-\bar{y}\}\{x(t)-\bar{x}\}dt$ is the functional correlation between x and y and $\sigma_y^2=\int_0^1\{y(t)-\bar{y}\}^2dt$ (and σ_x^2 is defined similarly) is the functional variance. Further, we have that $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$.

Several take-home points are in order. First, we see that defining empirical means, covariances and variances for functional data is fairly straightforward. Secondly, we see that a version of linear regression applied to functional data is identical in all salient respects to ordinary linear regression. Thirdly, I hope that you can start to see a pattern that multivariate regression (in vector and more general spaces) can be built up easily by regression through the origin. It's an interesting, though tedious, exercise to derive multivariate regression only allowing oneself access to regression through the origin. We'll find a more convenient derivation in the next chapter. However, it's oddly pleasant that so much of multivariable regression relies on this simple result.

Chapter 5

Least squares

In this chapter we develop least squares.

5.1 Basics

Watch this video before beginning.

Let $\mathbf{X}$ be a design matrix, notationally its elements and column vectors are:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = [\mathbf{x}_1 \dots \mathbf{x}_p].$$

We are assuming that $n \geq p$ and $\mathbf{X}$ is of full (column) rank. Consider ordinary least squares

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^t\mathbf{y} - 2\mathbf{y}^t\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}. \tag{5.1}$$

If we were to minimize (5.1) with respect to $\boldsymbol{\beta}$, consider using our matrix derivative results from Chapter 2.

$$\frac{d}{d\boldsymbol{\beta}}(5.1) = -2\mathbf{X}^t\mathbf{y} + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}.$$

Solving for 0 leads to the so called normal equations:

$$\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y}.$$

Recall that $\mathbf{X}^t\mathbf{X}$ retains the same rank as $\mathbf{X}$. Therefore, it is a full rank $p \times p$ matrix and hence is invertible. We can then solve the normal equations as:

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}. \tag{5.2}$$

The Hessian of (5.1) is simply $2\mathbf{X}^t\mathbf{X}$, which is positive definite. (This is clear since for any non-zero vector, $\mathbf{a}$, we have that $\mathbf{X}^t\mathbf{a}$ is non-zero since $\mathbf{X}$ is full rank and then $\mathbf{a}^t\mathbf{X}^t\mathbf{X}\mathbf{a} = ||\mathbf{X}\mathbf{a}||^2 > 0$.) Thus, the root of our derivative is indeed a minimum.

5.1.1 Coding example

Watch this video before beginning.

```
> y = swiss$Fertility
> x = as.matrix(swiss[,-1])
> solve(t(x) %*% x, t(x) %*% y)
                       [,1]
1                66.9151817
Agriculture      -0.1721140
Examination      -0.2580082
Education        -0.8709401
Catholic          0.1041153
Infant.Mortality  1.0770481
> summary(lm(y ~ x - 1))$coef
                   Estimate  Std. Error   t value     Pr(>|t|)
x1               66.9151817 10.70603759  6.250229 1.906051e-07
xAgriculture     -0.1721140  0.07030392 -2.448142 1.872715e-02
xExamination     -0.2580082  0.25387820 -1.016268 3.154617e-01
xEducation       -0.8709401  0.18302860 -4.758492 2.430605e-05
xCatholic         0.1041153  0.03525785  2.952969 5.190079e-03
xInfant.Mortality  1.0770481  0.38171965  2.821568 7.335715e-03
```

5.2 A second derivation

Watch this video before beginning.

If you know the answer first, it's possible to derive the minimum to the least squares equation without taking any derivatives.

$$
\begin{aligned}
||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 &= ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\beta} - \mathbf{X}\boldsymbol{\beta}|| \\
&= ||\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}||^2 + 2(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})^t(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}) + ||\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}||^2 \\
&\geq ||\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}||^2 + 2(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}) \\
&= ||\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}||^2 + 2(\mathbf{y} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y})^t\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 + 2\mathbf{y}^t(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)^t\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 + 2\mathbf{y}^t(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= ||\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}||^2 + 2\mathbf{y}^t(\mathbf{X} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= ||\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}||^2 + 2\mathbf{y}^t(\mathbf{X} - \mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= ||\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}||^2
\end{aligned}
$$

Thus, any value of β that we plug into the least squares equation is going to give us a larger norm than if we plug in $\hat{\beta}$ so that it is the unique minimum. Notice that going from line 5 to 6, we used the fact that $\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ is symmetric. (It is also idempotent.)

Also, we used a fact that is very useful in general, $\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ is orthogonal to any linear combination of the columns of $\mathbf{X}$. This is try since if $\mathbf{X}\mathbf{a}$ is such a combination, then

$$\{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\}bX\mathbf{a} = \{\mathbf{X} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\}\mathbf{a} = (\mathbf{X} - \mathbf{X})\mathbf{a} = 0.$$

This fact is extremely handy in working with linear models.

5.3 Connection with linear regression

Watch this video before beginning.

Recall that the slope from linear regression worked out to be

$$\langle \mathbf{x} - \bar{x}\mathbf{1}_n, \mathbf{x} - \bar{x}\mathbf{1}_n\rangle^{-1}\langle \mathbf{x} - \bar{x}\mathbf{1}_n, \mathbf{y} - \bar{y}\mathbf{1}_n\rangle = \hat{\sigma}_X^{-2}\hat{\sigma}_{XY}^2$$

where $\hat{\sigma}_{XY}^2$ is the empirical covariance between X and Y. (We rewrote this formula using the more convenient correlation.) In this form it is the covariance between x and y divided by the variance of the x's. Let's consider extending this in our matrix results.

Let $\mathbf{X} = [\mathbf{1}_n\mathbf{X}_1]$, thus $\mathbf{X}$ contains an intercept and then $p-1$ other regressors. Similarly let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^t = (\beta_0\beta_1^t)^t$. Consider now least squares

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|| = ||\mathbf{y} - \mathbf{1}_n\beta_0 - \mathbf{X}_1\boldsymbol{\beta}_1||$$

If we were to hold β_1 fixed we are faced with a mean only regression problem and the solution to $\beta_0(\boldsymbol{\beta}_1)$ is

$$\frac{1}{n}(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1)^t\mathbf{1}_n = \bar{y} - \bar{\mathbf{x}}^t\boldsymbol{\beta}_1$$

where $\bar{\mathbf{x}}$ is the columnwise means of $\mathbf{X}$. Plugging this back into our least squares equation for β_0 we get

$$||\mathbf{y} - \mathbf{1}_n\bar{y} - (\mathbf{X}_1 - \mathbf{1}_n\bar{\mathbf{x}}^t)\boldsymbol{\beta}_1||^2 = ||\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1||^2$$

where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ are the centered versions of $\mathbf{y}$ and $\mathbf{X}$. This is again just the least squares equation with the centered variables and thus we get that

$$\hat{\beta}_1 = (\tilde{\mathbf{X}}^t\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{Y}} = \hat{\beta}_1 = \left(\frac{1}{n-1}\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}\right)^{-1}\frac{1}{n-1}\tilde{\mathbf{X}}\tilde{\mathbf{Y}}.$$

The matrix $\frac{1}{n-1}\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$ is the empirical variance covariance matrix of the columns of $\mathbf{X}$ while $\frac{1}{n-1}\tilde{\mathbf{X}}\tilde{\mathbf{Y}}$ is the vector of correlations of $\mathbf{y}$ with the columns of $\mathbf{X}$. Therefore, if we include an intercept, our slope estimate is

$$\hat{\boldsymbol{\Sigma}}_{XX}^{-1}\hat{\boldsymbol{\rho}}_{XY}$$

the inverse of the variance matrix associated with $\mathbf{X}$ time the correlation matrix between $\mathbf{X}$ and $\mathbf{Y}$. This draws an exact parallel with the result from linear regression.

5.4 Projections

Watch this video before beginning.

The vector of fitted values is

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

and the vector of residuals is given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{y}.$$

Thus, multiplication by the matrix $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ takes any vector in $\mathbb{R}^n$ and produces the fitted values. The matrix $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ is called the projection matrix (for reasons that will become obvious) or the "hat matrix" (I guess because it transforms our Y into "Y hat", then perhaps it should be called the "hatting" matrix?).

Multiplication by $(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)$ produces the residuals. Notice that since the $\hat{\mathbf{y}}$ vector is a linear combination of the $\mathbf{X}$, it is orthogonal to the residuals:

$$\hat{\mathbf{y}}^t\mathbf{e} = \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{y} = 0.$$

It is useful to think of least squares in the terms of projections. Consider the column space of the design matrix, $\Gamma = \{\mathbf{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}$. This p dimensional space lives in $\mathbb{R}^n$, so think of a plane in $\mathbb{R}^3$. Consider the vector $\mathbf{y}$ which lives in $\mathbb{R}^n$. Multiplication by the matrix $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ projects $\mathbf{y}$ into Γ. That is,

$$\mathbf{y} \rightarrow \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

is the linear projection map between $\mathbb{R}^n$ and Γ. The point $\hat{\mathbf{y}}$ is the point in Γ that is closest to $\mathbf{y}$ and $\hat{\boldsymbol{\beta}}$ is the specific linear combination of the columns of $\mathbf{X}$ that yields $\hat{\mathbf{y}}$. $\mathbf{e}$ is the vector connecting $\mathbf{y}$ and $\hat{\mathbf{y}}$, and it is orthogonal to all elements in Γ.

Logically the projection matrix must be idempotent. Consider that for any vector, multiplication by the projection matrix finds the closest element in Γ. Therefore, we can't multiply again and find a closer one. That is, $Py = P^2y$ for a projection matrix P and any y and thus $P = P^2$. Since we are dealing in Euclidean spaces, a projection matrix must also be symmetric. To see this, note that the residual must be orthogonal to any projected point, $< (I - P)y, Pw >= 0 = y^t(I - P^t)Pw$. Since this holds for all y and w, it must be the case that $(I - P^t)P = 0$ or in other words $P = P^tP$. Since the right hand side is symmetric, P must be symmetric. It's worth noting that this result is dependent the use of the Euclidean metric. If the inner product is $< a, b >= a^t\Sigma^{-1}b$ (Mahalanobis distance), the the projection metric is necessarily idempotent, but not symmetric.

Thinking this helps us interpret statistical aspects of least squares. First, if $\mathbf{W}$ is any $p \times p$ invertible matrix, then the fitted values, $\hat{\mathbf{y}}$ will be the same for the design matrix $\mathbf{XW}$. This is because the spaces

$$\{\mathbf{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}$$

and

$$\{\mathbf{XW}\boldsymbol{\gamma} \mid \boldsymbol{\gamma} \in \mathbb{R}^p\}$$

are the same, since if $\mathbf{a} = \mathbf{X}\boldsymbol{\beta}$ then $\mathbf{a} = \mathbf{X}\boldsymbol{\gamma}$ via the relationship $\gamma = \mathbf{W}\boldsymbol{\beta}$ and thus any element of the first space is in the second. The same argument implies in the other direction, thus the two spaces are the same.

Therefore, any linear reorganization of the columns of $\mathbf{X}$ results in the same column space and thus the same fitted values. Furthermore, any addition of redundant columns to $\mathbf{X}$ adds nothing to the column space, and thus it's clear what the fit should be in the event that $\mathbf{X}$ is not full rank. Any full rank subset of the columns of $\mathbf{X}$ defines the same column and thus the same fitted values.

5.5 Full row rank case

In the case where $\mathbf{X}$ is $n \times n$ of full rank, then the columns of $\mathbf{X}$ form a basis for $\mathbb{R}^{\ltimes}$. In this case, $\hat{\mathbf{y}} = \mathbf{y}$, since $\mathbf{y}$ lives in the space spanned by the columns of $\mathbf{X}$. All the linear model accomplishes is a lossless linear reorganization of $\mathbf{y}$. This is perhaps surprisingly useful, especially when the columns of $\mathbf{X}$ are orthonormal ($\mathbf{X}^t\mathbf{X} = \mathbf{I}$). In this case, the function that takes the outcome vector and converts it to the coefficients is called a ”transform”. The most well known versions of transforms are Fourier and wavelet.

5.6 A third derivation

Watch this video before beginning.

In this section we generate a third derivation of least squares. For vectors $\mathbf{a}$ (outcome) and $\mathbf{b}$ (predictor), define the coefficient function as:

$$c(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{||\mathbf{b}||^2}.$$

and the residual function as

$$e(\mathbf{a}, \mathbf{b}) = \mathbf{a} - c(\mathbf{a}, \mathbf{b})\mathbf{b}$$

We argue that the least squares estimate of outcome $\mathbf{y}$ for predictor matrix $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_p]$ is obtained by taking successive residuals, in the following sense. Consider the least squares equation holding $\beta_2, \ldots, \beta_p$ fixed:

$$||\mathbf{y} - \mathbf{x}_1\boldsymbol{\beta}_1 - \ldots - \mathbf{x}_p\boldsymbol{\beta}_p||^2. \tag{5.3}$$

This is greater than or equal to if we replace β_1 by it's estimate with the remainder fixed. That estimate being:

$$c(\mathbf{y} - \mathbf{x}_2\boldsymbol{\beta}_2 - \ldots, \mathbf{x}_p\boldsymbol{\beta}_p, \mathbf{x}_1).$$

Plugging that back into the least squares equation we get

$$(5.3) \geq ||e(\mathbf{y}, \mathbf{x}_1) - e(\mathbf{x}_2, \mathbf{x}_1)\beta_2, \ldots, e(\mathbf{x}_p, \mathbf{x}_1)\beta_p||^2. \tag{5.4}$$

Thus we have a new least squares equation with all residuals having ”removed” $\mathbf{x}_1$ from all other regressors and the outcome. Then we can repeat this process again holding $\beta_3, \ldots, \beta_p$ fixed and obtain

$$(5.4) \geq ||e\{e(\mathbf{y}, \mathbf{x}_1), e(\mathbf{x}_2, \mathbf{x}_1)\} - e\{e(\mathbf{x}_3, \mathbf{x}_1), e(\mathbf{x}_2, \mathbf{x}_1)\}\beta_3, \ldots, e\{e(\mathbf{x}_p, \mathbf{x}_1), e(\mathbf{x}_2, \mathbf{x}_1)\}\beta_p||^2.$$

This could then be iterated to the p^{th} regressor. Moreover, because we know the same inequalities will be obtained no matter what order we get to the p^{th} regressor we can conclude that the order of taking residuals doesn't matter. Furthermore, picking the p^{th} coefficient was arbitrary as well, so the same conclusion applies to all regressors: the least squares estimate for all coefficients can be obtained by iteratively taking residuals with all of the other regressors (in any order).

This is interesting for many reasons. First, it is interesting to note that one need only regression through the origin to develop full multivariable regression. Secondly it helps us interpret our regression coefficients and how they are ”adjusted” for the other variables.

There was nothing in particular about using vectors. If $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$, two submatrices of size p_1 and p_2, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^t \boldsymbol{\beta}_2^t)^t$ consider minimizing

$$||\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2||^2.$$

If $\boldsymbol{\beta}_2$ were held fixed, this would be maximized at

$$\boldsymbol{\beta}_1(\boldsymbol{\beta}_2) = (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^2(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2).$$

Plugging that back in we obtain a smaller quantity

$$||\{\mathbf{I} - (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^2\}\mathbf{y} - \{\mathbf{I} - (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^2\}\mathbf{X}_2\boldsymbol{\beta}_2||^2$$

This is equivalent to the residual of $\mathbf{y}$ having regressed out $\mathbf{X}_1$ and the residual matrix of $\mathbf{X}_2$ having regressed $\mathbf{X}_1$ out of every column. Thus out β_2 estimate will be the regression matrix of these residuals. Again, this explains why $\boldsymbol{\beta}_2$'s estimate has been adjusted for $\mathbf{X}_1$, both the outcome and the $\mathbf{X}_2$ predictors have been orthogonalized to the space spanned by the columns of $\mathbf{X}_1$!

5.6.1 Coding example

Watch this video before beginning.

```
> y = swiss$Fertility
> x = as.matrix(swiss[,-1])
> x1 = x[,1 : 3]
> x2 = x[,4 : 6]
> solve(t(x) %*% x, t(x) %*% y)
                        [,1]
1                 66.9151817
Agriculture       -0.1721140
```

```
Examination      -0.2580082
Education        -0.8709401
Catholic          0.1041153
Infant.Mortality  1.0770481
> ey = y - x1 %*% solve(t(x1) %*% x1, t(x1) %*% y)
> ex2 = x2 - x1 %*% solve(t(x1) %*% x1) %*% t(x1) %*% x2
> solve(t(ex2) %*% ex2, t(ex2) %*% ey)
                       [,1]
Education        -0.8709401
Catholic          0.1041153
Infant.Mortality  1.0770481
```

Chapter 6

Conceptual examples of least squares

6.1 Mean only regression

Watch this video before beginning.

If our design matrix is $\mathbf{X} = \mathbf{1}_n$, we see that our coefficient estimate is:

$$(\mathbf{1}_n^t \mathbf{1}_n)^{-1} \mathbf{1}_n \mathbf{y} = \bar{y}.$$

6.2 Regression through the origin

If our design matrix is $\mathbf{X} = \mathbf{x}$, we see that our coefficient is

$$(\mathbf{x}^t \mathbf{x})^{-1} \mathbf{x}^t \mathbf{y} = \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{||\mathbf{x}||^2}.$$

6.3 Linear regression

Section 5.3 of the last chapter showed that multivariable least squares was the direct extension of linear regression (and hence reduces to it).

6.4 ANOVA

Watch this video before beginning.

Let $\mathbf{y} = [y_{11}, \ldots y_{JK}]$ and our design matrix look like

$$\mathbf{X} = \left[\begin{array}{cccc} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \ldots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \ldots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{array}\right] = \mathbf{I}_K \otimes \mathbf{1}_n,$$

where $\otimes$ is the Kronecker product. That is, our $\mathbf{y}$ arises out of J groups where there is K measurements per group. Let $\bar y_j$ be the mean of the $\mathbf{y}$ measurements in group j. Then

$$\mathbf{X}^t\mathbf{y} = \left[\begin{array}{c} K\bar y_1 \\ \vdots \\ K\bar y_J \end{array}\right],$$

Note also that $\mathbf{X}^t\mathbf{X} = k\mathbf{I}$. Therefore, $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = (\bar y_1 \ldots, \bar y_J)^t$. Thus, if our design matrix parcels $\mathbf{y}$ into groups, the coefficients are the group means.

Some completing thoughts on ANOVA.

6.5 ANCOVA

Watch this video before beginning.

Consider now an instance where

$$\mathbf{X} = \left[\begin{array}{ccc} 1 & 0 & x_{11} \\ 1 & 0 & x_{12} \\ \vdots & \ldots & \vdots \\ 1 & 0 & x_{1n} \\ 0 & 1 & x_{21} \\ 0 & 1 & x_{22} \\ \vdots & \ldots & \ldots \\ 0 & 1 & x_{2n} \end{array}\right] = [\mathbf{I}_2 \otimes \mathbf{1}_n \;\; \mathbf{x}].$$

That is we want to project $\mathbf{y}$ onto the space spanned by two groups and a regression variable. This is effectively fitting two parallel lines to the data. Let $\boldsymbol{\beta} = (\mu_1 \, \mu_2 \, \beta)^t = (\boldsymbol{\mu}^t \, \beta)^t$. Denote the outcome vector, $\mathbf{y}$, as comprised of y_{ij} for $i = 1, 2$ and $j = 1, \ldots, n$ stacked in the relevant order. Imagine holding β fixed.

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{y} - \mathbf{x}\beta - (\mathbf{I}_2 \otimes \mathbf{1}_n)\boldsymbol{\mu}||^2 \tag{6.1}$$

Then we are in the case of the previous section and the best estimate of $\boldsymbol{\mu}$ are the group means $\frac{1}{n}(\mathbf{I}_2 \otimes \mathbf{1}_n)^t(y - \mathbf{x}\boldsymbol{\beta}) = (\bar y_1 \; \bar y_2)^t - (\bar x_1 \; \bar x_2)^t \beta$ where $\bar y_i$ and $\bar x_i$ are the group means of $\mathbf{y}$ and $\mathbf{x}$ respectively. Then we have that (6.1) satisfies:

$$(6.1) \geq ||\mathbf{y} - \mathbf{x}\beta - (\mathbf{I}_2 \otimes \mathbf{1}_n)\{(\bar y_1 \; \bar y_2)^t - (\bar x_1 \; \bar x_2)^t \beta\}||^2 = ||\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\beta||^2$$

where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ are the group centered versions of $\mathbf{y}$ and $\mathbf{x}$. (That is $\tilde y_{ij} = y_{ij} - \bar y_i$, for example.) This is now regression through the origin yielding the solution

$$\hat\beta = \frac{\sum_{ij}(y_{ij} - \bar y_i)(x_{ij} - \bar x_i)}{\sum_{ij}(x_{ij} - \bar x_i)^2} = p\hat\beta_1 + (1-p)\hat\beta_2$$

where

$$p = \frac{\sum_j (x_{1j} - \bar x_1)^2}{\sum_{ij}(x_{ij} - \bar x_i)^2}$$

and

$$\hat\beta_i = \frac{\sum_j (y_{ij} - \bar y_i)(x_{ij} - \bar x_i)}{\sum_j (x_{ij} - \bar x_i)^2}.$$

That is, the estimated slope is a convex combination of the group-specific slopes weighted by the variability in the $\mathbf{x}$'s in the group. Furthermore, $\hat\mu_i = \bar y_i - \bar x_i \hat\beta$ and thus

$$\hat\mu_1 - \hat\mu_2 = (\bar y_1 - \bar y_2) - (\bar x_1 - \bar x_2)\hat\beta.$$

The ANCOVA model is extremely useful for visualizing adjustment in regression. See the video here for some examples.

Chapter 7

Bases

Recall that any set of linearly independent vectors forms a basis, specifically for the space spanned by linear combinations. Therefore, least squares is projecting our data into the space created by the basis defined by the columns of our design matrix. Of note, certain bases are of particular importance as the spaces that they create are contextually meaningful for many scientific problems. The most notable example are Fourier bases. In this case, we project our data into the space spanned by harmonic functions. In problem where the study of periodicities is of interest, this is of tremendous use.

7.1 Introduction to full rank bases

Watch this video before beginning.

When our $\mathbf{X}$ matrix is $n \times n$ of rank n, then the least squares fitted values $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{X}\boldsymbol{\beta} = \hat{\mathbf{y}}$ is simply a linear reorganization of $\mathbf{y}$ as it's projecting it from $\mathbb{R}^n$ to $\mathbb{R}^n$. Despite not summarizing $\mathbf{y}$ in any meaningful way, this is often a very meaningful thing to do, particularly when the basis is orthonormal. This full rank linear transformation of $\mathbf{y}$ is simply called a "transform". Notable bases then get named as their name then "transform". The best examples include the Fourier transform and the Wavelet transform. Often, because we're applying the transform to vectors with discrete indices, rather than continuous functions, the label "discrete" is affixed, such as the discrete Fourier transform. Looking back to our Hilbert space discussion from earlier the extension to continuous spaces is conceptually straightforward.

Let $\mathbf{X} = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_n]$ be our basis so that $\mathbf{X}^t\mathbf{X} = I$. In this case note that

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y} = \left(\begin{array}{c} \langle \mathbf{x}_1, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_n, \mathbf{y} \rangle \end{array}\right).$$

Thus, our coefficients are exactly the inner products of the basis elements (columns of $\mathbf{X}$) and the outcome. Our fitted values are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \langle \mathbf{x}_i, \mathbf{y} \rangle.$$

Consider deleting columns from $\mathbf{X}$, say

$$\mathbf{W} = [\mathbf{x}_{i_1} \ldots \mathbf{x}_{i_p}]$$

and minimizing $||\mathbf{y} - \mathbf{W}\boldsymbol{\gamma}||^2$. Since $\mathbf{W}$ is also orthonormal (but not full rank) we get that

$$\hat{\gamma} = \sum_{j=1}^{p} \langle \mathbf{x}_{i_j}, \mathbf{y} \rangle \quad \text{and} \quad \hat{\mathbf{y}} = \sum_{j=1}^{p} \mathbf{x}_{i_j} \langle \mathbf{x}_{i_j}, \mathbf{y} \rangle . = \sum_{j=1}^{p} \mathbf{x}_{i_j} \beta_{i_j}.$$

That is, the coefficients from the model with columns removed are identical to the coefficients from the full model. Transforms are often then accomplished by getting the full transform and using a subset of the coefficients to get the fitted values.

Watch this discussion of different kinds of bases.

7.2 Principal component bases

Watch this video before beginning.

One orthonormal basis is always at our disposal, the principal component basis. Consider the case where $\mathbf{X}$ has $p > n$ columns of full row rank. Let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^t$ be the singular value decomposition of $\mathbf{X}$ so that $\mathbf{U}$ is $n \times n$, so that, $\mathbf{U}^t\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^t$ is $n \times p$ so that $\mathbf{V}^t\mathbf{V} = \mathbf{I}$ and $\mathbf{D}$ is diagonal containing n singular values. The matrix $\mathbf{U}$ is a full rank version of the column space of $\mathbf{X}$. Notice that minimizing

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$$

has n equations and $p > n$ unknowns and that

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{y} - \mathbf{U}\mathbf{D}\mathbf{V}^t\boldsymbol{\beta}||^2$$

So that by defining $\boldsymbol{\gamma} = \mathbf{D}\mathbf{V}^t\boldsymbol{\beta}$ and minimizing

$$||\mathbf{y} - \mathbf{U}\boldsymbol{\gamma}||^2.$$

we have a full rank design matrix created out of the columns of $\mathbf{X}$ since $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{D}^{-1}$. The fitted values are merely $\mathbf{U}^t\mathbf{y}$.

Note, if $\mathbf{X}$ has been centered, then

$$\frac{1}{n-1}\mathbf{X}^t\mathbf{X} = \frac{1}{n-1}\mathbf{V}\mathbf{D}^2\mathbf{V}^t$$

is the covariance matrix between the columns of $\mathbf{X}$. Furthermore, notice that the total variability represented by the trace of the covariance matrix is,

$$\frac{1}{n-1}\text{tr}(\mathbf{X}^t\mathbf{X}) = \frac{1}{n-1}\text{tr}(\mathbf{V}\mathbf{D}^2\mathbf{V}^t) = \frac{1}{n-1}\text{tr}(\mathbf{D}^2\mathbf{V}^t\mathbf{V}) = \frac{1}{n-1}\text{tr}(\mathbf{D}^2)$$

The sum of the squared singular values equals the total variability in the design matrix. The singular vectors and values are typically ordered in the terms of decreasing variability.

Therefore, keeping only a few of them represents a dimension reduction that preserves the greatest amount of variability.

Thus, we once one calculates $\mathbf{U}^t\mathbf{Y}$ we have all possible submodel fits of the columns of $\mathbf{U}$, where $\mathbf{U}$ is an meaningful summary of $\mathbf{X}$. Typically one takes a few of the first columns of $\mathbf{U}$ so that the related eigenvalues explain a large proportion of the total variability. We haven't discussed an intercept. However, one usually mean centers $\mathbf{Y}$ and $\mathbf{X}$ first.

7.2.1 Coding example

Watch this video before beginning.

The following code goes through calculation of SVD and eigenvalue decompositions.

```
data(swiss)
y = swiss$Fertility
x = as.matrix(swiss[,-1])
n = nrow(x)
decomp = princomp(x, cor = TRUE)
plot(cumsum(decomp$sdev^2) / sum(decomp$sdev^2), type = "l")
decomp2 = eigen(cor(x))
xnorm = apply(x, 2, function(z) (z - mean(z)) / sd(z))
decomp3 = svd(xnorm)
round(rbind(decomp2$vectors, decomp$loadings, decomp3$v),3)
round(rbind(decomp2$values, decomp$sdev^2, decomp3$d ^ 2 / (n - 1)), 3)
plot(decomp3$u[,1], decomp$scores[,1])
plot(decomp3$u[,1], xnorm %*% decomp2$vectors %*% diag(1 / sqrt(decomp2$values))[,1])
```

Chapter 8

Residuals and variability

8.1 Residuals

Watch this video before beginning.

The residuals are the variability left unexplained by the projection onto the linear space spanned by the design matrix. The residuals are othogonal to the space spanned by the design matrix and thus are othogonal to the design matrix itself.

We define the residuals as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}.$$

Thus, our least squares solution can be though of as minimizing the squared norm of the residuals. Notice further that by expanding the column space of $\mathbf{X}$ by adding any new linearly indpendent variables, the normal of the residuals must decrease. In other words, if we add any non-redundant regressors, we necessarily remove residual variability. Furthermore, as we already know, $\mathbf{X}$ is of full column rank, then our residuals are all zero, since $\mathbf{y} = \hat{\mathbf{y}}$.

Notice that the residuals are equal to:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\}\mathbf{y}.$$

Thus multiplication by the matrix $\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ transforms a vector to the residual. This matrix is interesting for several reasons. First, note that $\{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\}\mathbf{X} = 0$ thus making the residuals orthogonal to any vector, $\mathbf{X}\gamma$, in the space spanned by the columns of $\mathbf{X}$. Secondly, it is both symmetric and idempotent.

A consequence of the orthogonality is that if an intercept is included in the model, the residuals sum to 0. Specifically, since the residuals are orthogonal to any column of $\mathbf{X}$, $\mathbf{e}^t\mathbf{1} = 0$.

8.2 Partitioning variability

Watch this video before beginning.

For convenience, define $\mathbf{H_X} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$. Note that the variability in a vector $\mathbf{y}$ is estimated by

$$\frac{1}{n-1}\mathbf{y}^t(\mathbf{I} - \mathbf{H_1})\mathbf{y}.$$

Omitting the $n-1$ term define the total sums of squares as

$$\text{SS}_{Tot} = ||\mathbf{y} - \bar{y}\mathbf{1}||^2 = \mathbf{y}^t(\mathbf{I} - \mathbf{H_1})\mathbf{y}.$$

This is an unscaled measure of the total variability in the sample. Given a design matrix, $\mathbf{X}$, define the residual sums of squares as

$$\text{SS}_{Res} = ||\mathbf{y} - \hat{\mathbf{y}}||^2 = \mathbf{y}^t(\mathbf{I} - \mathbf{H_X})\mathbf{y}$$

and the regression sums of squares as

$$\text{SS}_{Reg} = ||\bar{Y}\mathbf{1} - \hat{\mathbf{y}}||^2 = \mathbf{y}^t(\mathbf{H_X} - \mathbf{H_1})\mathbf{y}.$$

The latter equality is obtained by the following. First note that since $(\mathbf{I} - \mathbf{H_X})\mathbf{1} = 0$ (since $\mathbf{X}$ contains an intercept) we have that $\mathbf{H_X}\mathbf{1} = \mathbf{1}$ and then $\mathbf{H_X}\mathbf{H_1} = \mathbf{H_1}$ and $\mathbf{H_1} = \mathbf{H_1}\mathbf{H_X}$. Also, note that $\mathbf{H_X}$ is symmetric and idempotent. Now we can perform the following manipulation

$$\begin{aligned}
||\bar{Y}\mathbf{1} - \hat{\mathbf{y}}||^2 &= \mathbf{y}^t(\mathbf{H_X} - \mathbf{H_1})^t(\mathbf{H_X} - \mathbf{H_1})\mathbf{y} \\
&= \mathbf{y}^t(\mathbf{H_X} - \mathbf{H_1})(\mathbf{H_X} - \mathbf{H_1})\mathbf{y} \\
&= \mathbf{y}^t(\mathbf{H_X} - \mathbf{H_1}\mathbf{H_X} - \mathbf{H_X}\mathbf{H_1} + \mathbf{H_1})\mathbf{y} \\
&= \mathbf{y}^t(\mathbf{H_X} - \mathbf{H_1})\mathbf{y}.
\end{aligned}$$

Using this identity we can now show that

$$\begin{aligned}
\text{SS}_{Tot} &= \mathbf{y}^t(\mathbf{I} - \mathbf{H_1})\mathbf{y} \\
&= \mathbf{y}^t(\mathbf{I} - \mathbf{H_X} + \mathbf{H_X} - \mathbf{H_1})\mathbf{y} \\
&= \mathbf{y}^t(\mathbf{I} - \mathbf{H_X})\mathbf{y} + \mathbf{y}^t(\mathbf{H_X} - \mathbf{H_1})\mathbf{y} \\
&= \text{SS}_{Res} + \text{SS}_{Reg}
\end{aligned}$$

Thus our total sum of squares partitions into the residual and regression sums of squares. We define

$$R^2 = \frac{\text{SS}_{Reg}}{\text{SS}_{Tot}}.$$

as the percentage of our total variability explained by our model. Via our equality above, this is guaranteed to be between 0 and 1.

Chapter 9

Expectations

Up to this point, our exploration of linear models only relied on least squares and projections. We begin now discussing the statistical properties of our estimators. We start by defining expected values. We assume that the reader has basic univariate mathematical statistics.

9.1 Expected values

Watch this video before beginning.

If X is a random variable having density funciton f, the k^{th} moment is defined as

$$E[X] = \int_{-\infty}^{\infty} x^k f(x) dx.$$

In the multivariate case where $\mathbf{X}$ is a random vector then the k^th moment of element i of the vector is given by

$$E[X_i^k] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_i^k f(x_1, \ldots, x_n) dx_1, \ldots, dx_n.$$

It is worth asking if this definition is consistent with all of the subdistributions defined by the subvectors of $\mathbf{X}$. Let $i_1, \ldots, i_p$ is any subset of indices of $1, \ldots, n$ and $i_{p+1}, \ldots, i_n$ are the remaining, then the joint distribution of $(X_{i_1}, \ldots, X_{i_p})^t$ is

$$g(x_{i_1}, \ldots, x_{i_p}) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_{i_{p+1}}, \ldots, dx_{i_n}.$$

The k^{th} moment of X_{i_j} for $j \in \{1, \ldots, p\}$ is equivalently:

$$\begin{aligned} E[X_{i_j}] &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_{i_j}^k g(x_{i_1}, \ldots, x_{i_p}) dx_{i_1}, \ldots, d_{x_{ip}} \\ &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_{i_j}^k f(x_1, \ldots, x_n) dx_1, \ldots, dx_n. \end{aligned}$$

(HW, prove this.) Thus, if we know only the marginal distribution of X_{i_j} or any level of joint information, the expected value is the same.

If $\mathbf{X}$ is any random vector or matrix, the $E[\mathbf{X}]$ is simply the elementwise expected value defined above. Often we will write $E[\mathbf{X}] = \boldsymbol{\mu}$, or some other Greek letter, adopting the notation that population parameters are Greek. Standard notation is hindered somewhat in that uppercase letters are typically used for random values, though are also used for matrices. We hope that the context will eliminate confusion.

Watch this video before beginning.

Expected value rules translate well in the multivariate settings. If $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are vectors or matrices that satisfy the operations then

$$E[\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{Y} + \mathbf{C}] = \mathbf{A}E[\mathbf{X}] + \mathbf{B}E[\mathbf{Y}] + \mathbf{C}.$$

Further, expected values commute with transposes and traces

$$E[\mathbf{X}^t] = E[\mathbf{X}]^t$$

and

$$E[\text{tr}(\mathbf{X})] = \text{tr}(E[\mathbf{X}]).$$

9.2 Variance

Watch this video before beginning.

The multivariate variance of random vector $\mathbf{X}$ is defined as

$$\text{Var}(\mathbf{X}) = \boldsymbol{\Sigma} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^t].$$

Direct use of our matrix rules for expected values gives us the analog of the univariate shortcut formula

$$\boldsymbol{\Sigma} = E[\mathbf{X}\mathbf{X}^t] - \boldsymbol{\mu}\boldsymbol{\mu}^t.$$

Variance satisfy the properties

$$\text{Var}(\mathbf{A}\mathbf{X} + \mathbf{B}) = \mathbf{A}\text{Var}(\mathbf{X})\mathbf{A}^t.$$

9.3 Multivariate covariances

Watch this video before beginning.

The multivariate covariance is given by

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{Y} - \boldsymbol{\mu}_y)^t] = E[\mathbf{X}\mathbf{Y}^t] - \boldsymbol{\mu}_x\boldsymbol{\mu}_y^t.$$

This definition applies even if $\mathbf{X}$ and $\mathbf{Y}$ are of different length. Notice the multivariate covariance is not symmetric in its arguments. Moreover,

$$\text{Cov}(\mathbf{X}, \mathbf{X}) = \text{Var}(\mathbf{X}).$$

Covariances satisfy some useful rules in that

$$\text{Cov}(\mathbf{AX}, \mathbf{BY}) = \mathbf{A}\text{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^t$$

and

$$\text{Cov}(\mathbf{X} + \mathbf{Y}, \mathbf{Z}) = \text{Cov}(\mathbf{X}, \mathbf{Y}) + \text{Cov}(\mathbf{X}, \mathbf{Z})$$

Multivariate covariances are useful for sums of random vectors.

$$\text{Var}(\mathbf{X} + \mathbf{Y}) = \text{Var}(\mathbf{X}) + \text{Var}(\mathbf{Y}) + \text{Cov}(\mathbf{X}, \mathbf{Y}) + \text{Cov}(\mathbf{Y}, \mathbf{X}).$$

A nifty fact from covariances is that the covariance of $\mathbf{AX}$ and $\mathbf{BX}$ is $\mathbf{A\Sigma B}^t$. Thus $\mathbf{AX}$ and $\mathbf{BX}$ are uncorrelated iff $\mathbf{A\Sigma B}^t = \mathbf{0}$.

9.4 Quadratic form moments

Watch this video before beginning.

Let $\mathbf{X}$ be from a distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. Then

$$E[\mathbf{X}^t\mathbf{AX}] = \boldsymbol{\mu}^t\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{A\Sigma}).$$

Proof

$$\begin{aligned}
E[\mathbf{X}^t\mathbf{AX}] &= E[\text{tr}(\mathbf{X}^t\mathbf{AX})] \\
&= E[\text{tr}(\mathbf{AXX}^t)] \\
&= \text{tr}(E[\mathbf{AXX}^t]) \\
&= \text{tr}(\mathbf{A}E[\mathbf{XX}^t]) \\
&= \text{tr}\{\mathbf{A}[\text{Var}(\mathbf{X}) + \boldsymbol{\mu\mu}^t]\} \\
&= \text{tr}\{\mathbf{A\Sigma} + \mathbf{A}\boldsymbol{\mu\mu}^t\} \\
&= \text{tr}(\mathbf{A\Sigma}) + \text{tr}(\mathbf{A}\boldsymbol{\mu\mu}^t) \\
&= \text{tr}(\mathbf{A\Sigma}) + \text{tr}(\boldsymbol{\mu}^t\mathbf{A}\boldsymbol{\mu}) \\
&= \text{tr}(\mathbf{A\Sigma}) + \boldsymbol{\mu}^t\mathbf{A}\boldsymbol{\mu}
\end{aligned}$$

9.5 BLUE

Watch this video before beginning.

Now that we have moments, we can discuss mean and variance properties of the least squares estimators. Particularly, note that if Y satisfies $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ and $\text{Var}(Y) = \sigma^2\mathbf{I}$ then, $\hat{\boldsymbol{\beta}}$ satisfies:

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t E[\mathbf{Y}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Thus, under these conditions $\hat{\boldsymbol{\beta}}$ is unbiased. In addition, we have that

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}\{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}\} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\text{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = (\mathbf{X}^t\mathbf{X})^{-1}\sigma^2.$$

We can extend these results to linear contrasts of $\boldsymbol{\beta}$ to say that $\mathbf{q}^t\hat{\boldsymbol{\beta}}$ is the *best* estimator of $\mathbf{q}^t\boldsymbol{\beta}$ in the sense of minimizing the variance among linear (in $\mathbf{Y}$) unbiased estimators. It is important to consider unbiased estimators, since we could always minimize the variance by defining an estimator to be constant (hence variance 0). If one removes the restriction of unbiasedness, then minimum variance cannot be the definition of "best". Often one then looks to mean squared error, the squared bias plust the variance, instead. In what follows we only consider linear unbiased estimators.

We give Best Linear Unbiased Estimators the acronym BLUE. It is remarkable easy to prove the result.

Consider estimating $\mathbf{q}^t\boldsymbol{\beta}$. Clearly, $\mathbf{q}^t\hat{\boldsymbol{\beta}}$ is both unbiased and linear in $\mathbf{Y}$. Also note that Var$(\mathbf{q}^t\hat{\boldsymbol{\beta}}) = \mathbf{q}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{q}\sigma^2$. Let $\mathbf{k}^t\mathbf{Y}$ be another linear unbiased estimator, so that $E[\mathbf{k}^t\mathbf{Y}] = \mathbf{q}^t\boldsymbol{\beta}$. But, $E[\mathbf{k}^t\mathbf{Y}] = \mathbf{k}^t\mathbf{X}\boldsymbol{\beta}$. It follows that since $\mathbf{q}^t\boldsymbol{\beta} = \mathbf{k}^t\mathbf{X}\boldsymbol{\beta}$ must hold for all possible $\boldsymbol{\beta}$, we have that $\mathbf{k}^t\mathbf{X} = \mathbf{q}^t$. Finally note that

$$\text{Cov}(\mathbf{q}^t\hat{\boldsymbol{\beta}}, \mathbf{k}^t\mathbf{Y}) = \mathbf{q}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{k}^t\sigma^2.$$

Since $\mathbf{k}^t\mathbf{X} = \mathbf{q}^t$, we have that

$$\text{Cov}(\mathbf{q}^t\hat{\boldsymbol{\beta}}, \mathbf{k}^t\mathbf{Y}) = \text{Var}(\mathbf{q}^t\boldsymbol{\beta}).$$

Now we can execute the proof easily.

$$\begin{aligned}\text{Var}(\mathbf{q}^t\hat{\boldsymbol{\beta}} - \mathbf{k}^t\mathbf{Y}) &= \text{Var}(\mathbf{q}^t\hat{\boldsymbol{\beta}}) + \text{Var}(\mathbf{k}^t\mathbf{Y}) - 2\text{Cov}(\mathbf{q}^t\hat{\boldsymbol{\beta}}, \mathbf{k}^t\mathbf{Y}) \\ &= \text{Var}(\mathbf{k}^t\mathbf{Y}) - \text{Var}(\mathbf{q}^t\hat{\boldsymbol{\beta}}) \\ &\geq 0.\end{aligned}$$

Here the final inequality arises as variances have to be non-negative. Then we have that Var$(\mathbf{k}^t\mathbf{Y}) \geq$ Var$(\mathbf{q}^t\hat{\boldsymbol{\beta}})$ proving the result.

Notice, normality was not required at any point in the proof, only restrictions on the first two moments. In what follows, we'll see the consequences of assuming normality.

Chapter 10

The normal distribution

Watch this before beginning.

10.1 The univariate normal distribution

Z follows a standard normal distribution if its density is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2).$$

We write the associated distribution function as Φ. A standard normal variate has mean 0 and variance 1. All odd numbered moments are 0. The non-standard normal variate, say X, having mean μ and standard deviation σ can be obtained as $X = \mu + \sigma Z$. Conversely, $(X - \mu)/\sigma$ is standard normal if X is any non-standard normal. The non-standard normal density is:

$$\phi\left(\frac{x-\mu}{\sigma}\right)/\sigma$$

with distribution function $\Phi\left(\frac{x-\mu}{\sigma}\right)$.

10.2 The multivariate normal distribution

The multivariate standard normal distribution for a random vector $\mathbf{Z}$ has density given by:

$$(2\pi)^{-n/2} \exp(-||\mathbf{Z}||^2/2).$$

$\mathbf{Z}$ has mean $\mathbf{0}$ and variance $\mathbf{I}$. Non standard normal variates, say $\mathbf{X}$, can be obtained as $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Z}$ where $E[\mathbf{X}] = \boldsymbol{\mu}$ and Var$(X) = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$ (assumed to be positive definite). Conversely, one can go backwards with $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. The non-standard multivariate normal distribution is given by

$$(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\right\}.$$

Commit this density to memory.

The normal distribution is nice to work with in that all full row rank linear transformations of the normal are also normal. That is, if $\mathbf{a} + \mathbf{A}\mathbf{X}$ is normal if $\mathbf{A}$ is full row rank. Also, all conditional and submarginal distributions of the multivariate normal are also normal. (We'll discuss the conditional distribution more later.)

10.3 Singular normal

Watch this video before beginning.

What happens if the $\mathbf{A}$ in the paragraph above is not of full row rank? Then Var$(X) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t$ is not full rank. There are redundant elements of the vector $\mathbf{X}$ in that if you know some of them, you know the remainder. An example is our residuals. The matrix $(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)$ is not of full rank (it's rank is $n - p$). For example, if we include an intercept, the residuals must sum to 0. Know any $n-1$ of them and you know the n^{th}. A contingency for this is to define the singular normal distribution. A singular normal random variable is any random variable that can be written as $\mathbf{A}\mathbf{Z} + \mathbf{b}$ for a matrix $\mathbf{A}$ and vector $\mathbf{b}$ and standard normal vector $\mathbf{Z}$.

As an example, consider the case where $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Then the residuals, defined as $\{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\}\mathbf{Y} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\}(\mathbf{X}\boldsymbol{\beta} + \frac{1}{\sigma}\mathbf{Z})$ are a linear transformation of iid normals. Thus the residuals are singular normal.

The singular normal is such that all linear combinations and all submarginal and conditional distributions are also singular normal (prove this using the definition above!). The singular normal doesn't necessarily have a density function, because of the possibility of redundant entries. For example, the vector $(Z\ Z)$, where Z is a standard normal, doesn't have a joint density since the covariance matrix is $\mathbf{1}_{2\times 2}$, which isn't invertible.

In our treatment, the multivariate normal is the special case of the singular normal where the covariance matrix is full rank. In other treatments of linear models, the definition of the multivariate normal allows for the possibility of rank deficient covariance matrices. However, personally, I think the distinction is useful, so reserve the term multivariate normal for the full rank case.

10.4 Normal likelihood

Watch this video before beginning.

Let $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ then note that minus twice the log-likelihood is:

$$n\log(\sigma^2) + ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2/2\sigma^2$$

Holding σ^2 fixed we see that minmizing minus twice the log likelihood (thus maximizing the likelihood) yields the least squares solution:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}.$$

Since this doesn't depend on σ it is the MLE. Taking derivatives and setting equal to zero we see that

$$\hat \sigma^2 = ||\mathbf{e}||^2 / n$$

(i.e. the average of the squared residuals). We'll find that there's a potentially preferable unbiased estimate given by

$$S^2 = ||\mathbf{e}||^2/(n-p).$$

This model can be written in a likelihood equivalent fashion of

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$. However, one must be careful with specifying linear models this way. For example, if one wants to simulate $Y = X + Z$ where X and Z are generated independently, one can not equivalently simulate X by generating Y and Z independently and taking $Z - Y$. (Note Y and Z are correlated in the original simulation specification.) Writing out the distributions explicitly removes all doubt. Thus the linear notation, especially when there are random effects, is sort of lazy and imprecise (though everyone, your author included, uses it).

Let's consider another case, suppose that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are iid p vectors $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, disregarding constants, minus twice the log likelihood is

$$n \log |\boldsymbol{\Sigma}| + \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}).$$

Assume that $\boldsymbol{\Sigma}$ is known, then using our derivative rules from earlier, we can minimize this to obtain the MLE for $\boldsymbol{\mu}$

$$\hat \mu = \bar{\mathbf{Y}}$$

and the following for $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^t$$

Consider yet another case $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with known $\boldsymbol{\Sigma}$. Minus twice the log-likelihood is:

$$\log |\boldsymbol{\Sigma}| + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Using our matrix rules we find that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{y}.$$

This is the so-called weighted least squares estimate.

10.5 Conditional distributions

Watch this video before beginning.

The conditional distribution of a normal is of interest. Let $\mathbf{X} = [\mathbf{X}_1^t \ \mathbf{X}_2^t]^t$ be comprised of an $n_1 \times 1$ and $n_2 \times 1$ matrix where $n_1 + n_2 = n$. Assume that $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^t \ \boldsymbol{\mu}_2^t]$ and

$$\left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^t & \boldsymbol{\Sigma}_{22} \end{array}\right].$$

Consider now the conditional distribution of $\mathbf{X}_1 \mid \mathbf{X}_2$. A clever way to derive this (shown to me by a student in my class) is as follows let $\mathbf{Z} = \mathbf{X}_1 + \mathbf{A}\mathbf{X}_2$ where $\mathbf{A} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$. Then note that the covariance between $\mathbf{X}_2$ and $\mathbf{Z}$ is zero (HW).

Thus the distribution of $\mathbf{Z} \mid \mathbf{X}_2$ is equal to the distribution of $\mathbf{Z}$ and that it is normally distributed being a linear transformation of normal variates. Thus we know both

$$E[\mathbf{Z} \mid \mathbf{X}_2 = \mathbf{x}_2] = E[\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2] + \mathbf{A}E[\mathbf{X}_2 \mid \mathbf{X}_2 = \mathbf{x}_2] = E[\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2] + \mathbf{A}\mathbf{x}_2$$

and

$$E[\mathbf{Z} \mid \mathbf{X}_2 = \mathbf{x}_2] = E[\mathbf{Z}] = \boldsymbol{\mu}_1 + \mathbf{A}\boldsymbol{\mu}_2.$$

Setting these equal we get that

$$E[\mathbf{X}_1 \mid \mathbf{X}_2] = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2).$$

As a homework, using the same technique to derive the conditional variance

$$\text{Var}(\mathbf{Z} \mid \mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t.$$

10.5.1 Important example

Consider an example. Consider the vector $(\mathbf{Y} \ \mathbf{X}^t)^t$ where $\mathbf{Y}$ is 1×1 and $\mathbf{X}$ is $p \times 1$. Assume that the vector is normal with $E[\mathbf{Y}] = \mu_y$, $E[\mathbf{X}] = \boldsymbol{\mu}_x$ and the variances as σ_y^2 (1×1) and $\boldsymbol{\Sigma}_x$ $(p \times p)$ and covariance $\boldsymbol{\rho}_{xy}$ $(p \times 1)$.

Consider now predicting $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Clearly the a good estimate for this would be $E[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$. Our results suggest that $\mathbf{Y} \mid \mathbf{X} = \mathbf{x}$ is normal with mean:

$$\mu_y + \boldsymbol{\rho}_{xy}^t\boldsymbol{\Sigma}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) = \mu_y - \boldsymbol{\mu}_x\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\rho}_{xy} + \mathbf{x}^t\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\rho}_{xy} = \beta_0 + \mathbf{x}^t\boldsymbol{\beta}$$

where $\beta_0 = \mu_y - \boldsymbol{\mu}_x\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\rho}_{xy}$ and $\beta = \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\rho}_{xy}$. That is, the conditional mean in this case mirrors the linear model. The slope is defined exactly as the inverse of the variance/covariance matrix of the predictors times the cross correlations between the predictors and the response. We discussed the empirical version of this in Section 5.3 where we saw that the empirical coefficients are the inverse of the empirical variance of the predictors times the empirical correlations between the predictors and response. A similar mirroring occurs for the intercept as well.

This correspondence simply says that empirical linear model estimates mirror the population parameters if both the predictors and response are jointly normal. It also yields a motivation for the linear model in some cases where the joint normality of the predictor and response is conceptually reasonable. Though we note that often such joint normality is not reasonable, such as when the predictors are binary, even though the linear model remains well justified.

10.5.2 Gaussian graphical models

Consider our partitioned variance matrix.

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}12^t & \boldsymbol{\Sigma}_{22} \end{array}\right].$$

The upper diagonal element of $\boldsymbol{\Sigma}^{-1}$ is given by the inverse of $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t$. Recall that $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t = \text{Var}(\mathbf{X}_1 \mid \mathbf{X}_2)$. Suppose that $\mathbf{X}_1 = (X_{11}\ X_{12})^t$. Then this result suggests that X_{11} is independent of X_{12} given $\mathbf{X}_2$ if the $(1,2)$ off diagonal element of $\boldsymbol{\Sigma}^{-1}$ is zero. (Recall that independence and absence of correlation are equivalent in the multivariate normal.) There's nothing in particular about the first two positions, so we arrive at the following remarkable fact: whether or not the off diagonal elements of $\boldsymbol{\Sigma}^{-1}$ are zero determines the conditional independence of those random variables given the remainder. This forms the basis of so-called Gaussian graphical models. The graph defined by ascertaining which elements of $\boldsymbol{\Sigma}^{-1}$ are zero is called a conditional independence graph.

10.5.3 Bayes calculations

We assume a slight familiarity of Bayesian calculations and inference for this section. In a Bayesian analysis, one multiplies the likelihood times a prior distribution on the parameters to obtain a posterior. The posterior distribution is then used for inference. Let's go through a simple example. Suppose that $\mathbf{y} \mid \mu \sim N(\mu\mathbf{1}_n, \sigma^2 I)$ and $\mu \mid N(\mu_0, \tau^2)$ where $\mathbf{y}$ is $n \times 1$ and μ is a scalar. The normal distribution placed on μ is called the "prior" and μ_0 and τ^2 are assumed to be known. For this example, let's assume that σ^2 is also known. The goal is to calculate $\mu \mid \mathbf{y}$, the posterior distribution. This is done by multiplying prior times likelihood. Symbolically,

$$f(\text{Param}|\text{Data}) = \frac{f(\text{Param, Data})}{f(\text{Data})} \propto f(\text{Data}|\text{Param})f(\text{Param}) = \text{Likelihood} \times \text{Prior}.$$

Here, the proportional symbol, $\propto$, is with respect to the parameter.

Consider our problem, retaining only terms involving μ we have that minus twice the natural log of the distribution of $\mu \mid \mathbf{y}$ is given by

$$\begin{aligned} & -2\log(f(\mathbf{y} \mid \mu)) - 2\log(f(\mu)) \\ = \;& ||\mathbf{y} - \mu\mathbf{1}_n||^2/\sigma^2 + (\mu - \mu_0)^2/\tau^2 \\ = \;& -2\mu n\bar{y}/\sigma^2 + \mu^2 n/\sigma^2 + \mu^2/\tau^2 - 2\mu\mu_0/\tau^2 \\ = \;& -2\mu\left(\frac{\bar{y}}{\sigma^2/n} + \frac{\mu_0}{\tau^2}\right) + \mu^2\left(\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}\right) \end{aligned}$$

This is recognized as minus twice the log density of a normal distribution for μ with variance of

$$\text{Var}(\mu \mid \mathbf{y}) = \left(\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}\right)^{-1} = \frac{\tau^2\sigma^2/n}{\sigma^2/n + \tau^2}$$

and mean of

$$E[\mu \mid \mathbf{y}] = \left(\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}\right)^{-1} \left(\frac{\bar{y}}{\sigma^2/n} + \frac{\mu_0}{\tau^2}\right) = p\bar{y} + (1-p)\mu_0$$

where

$$p = \frac{\tau^2}{\tau^2 + \sigma^2/n}.$$

Thus $E[\mu \mid \mathbf{y}]$ is a mixture of the empirical mean and the prior mean. How much the means are weighted depends on the ratio of the variance of the mean (σ^2/n) and the prior variance (τ^2). As we collect more data ($n \to \infty$), or if the data is not noisy ($\sigma \to 0$) or we have a lot of prior uncertainty ($\tau \to \infty$) the empirical mean dominates. In contrast as we become more certain a priori ($\tau \to 0$) the prior mean dominates.

Chapter 11

Distributional results

In this chapter we assume that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. This is the standard normal linear model. We saw in the previous chapter that the maximum likelihood estimate for

11.1 Quadratic forms

Watch this video before beginning.

Let $\mathbf{A}$ be a symmetric matrix (not necessarily full rank) and let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $(\mathbf{X}-\boldsymbol{\mu})^t\mathbf{A}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2_p$ if $\mathbf{A}\boldsymbol{\Sigma}$ is idempotent and $p =$ Rank$(\mathbf{A})$.

As an example, note that $(\mathbf{X}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$ is clearly Chi-squared(n). This is most easily seen by the fact that $\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}-\boldsymbol{\mu}) = \mathbf{Z}$ is a vector of iid standard normals and thus the quadratic form is the sum of their squares. Using our result, $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ in this case and $\mathbf{A}\boldsymbol{\Sigma} = \mathbf{I}$, which is idempotent. The rank of $\boldsymbol{\Sigma}^{-1}$ is n.

Let's prove our result. Let $\mathbf{A} = \mathbf{V}\mathbf{D}^2\mathbf{V}^t$ where $\mathbf{D}$ is diagonal (with p non zero entries) and $\mathbf{V}\mathbf{V}^t = \mathbf{I}$. The assumption of idempotency gives us that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Sigma}$. Plugging in our decomposition for $\mathbf{A}$ and using the orthonormality of the columns of $\mathbf{V}$ we get that $\mathbf{D}\mathbf{V}^t\boldsymbol{\Sigma}\mathbf{V}\mathbf{D} = \mathbf{I}$. Then note that

$$(\mathbf{X}-\boldsymbol{\mu})^t\mathbf{A}(\mathbf{X}-\boldsymbol{\mu}) = (\mathbf{X}-\boldsymbol{\mu})^t\mathbf{V}\mathbf{D}\mathbf{D}\mathbf{V}^t(\mathbf{X}-\boldsymbol{\mu}). \tag{11.1}$$

But $\mathbf{D}\mathbf{V}^t(\mathbf{X}-\boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{D}\mathbf{V}^t\boldsymbol{\Sigma}\mathbf{V}\mathbf{D})$, which has variance equal to $\mathbf{I}$. Thus in Equation (11.1) we have the sum of p squared iid standard normals and is thus Chi-squared p.

11.2 Statistical properties

For homework, show that $\hat{\boldsymbol{\beta}}$ is normally distributed with moments:

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \quad \text{and} \quad \text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^t\mathbf{X})^{-1}\sigma^2.$$

The residual variance estimate is $S^2 = \frac{1}{n-p}\mathbf{e}^t\mathbf{e}$. Using Section 9.4 we see that it is unbiased, $E[S^2] = \sigma^2$. Note also that:

$$\begin{aligned}\frac{n-p}{\sigma^2}S^2 &= \frac{1}{\sigma^2}\mathbf{y}^t(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{y} \\ &= \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t\{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

is a quadratic form of as discussed in Section (11.1). Furthermore

$$\frac{1}{\sigma^2}\{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\}\text{Var}(Y) = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\},$$

which is idempotent. For symmetric idempotent matrices, the rank equals the trace; the latter of which is easily calculated as

$$\text{tr}\{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\} = \text{tr}\{\mathbf{I}\} - \text{tr}\{\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\} = n - \text{tr}\{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\} = n - p.$$

Thus, $\frac{n-p}{\sigma^2}S^2$ is Chi-squared $n - p$. The special case of this where $\mathbf{X}$ has only an intercept yields the usual empirical variance estimate.

11.2.1 Confidence interval for the variance

We can use the Chi-squared result to develop a confidence interval for the variance. Let $\chi^2_{n-p,\alpha}$ be the α quantile from the chi squared distribution with $n - p$ degrees of freedom. Then inverting the probability statement

$$1 - \alpha = P\left(\chi^2_{n-p,\alpha/2} \leq \frac{\mathbf{e}'\mathbf{e}}{n-p} \leq \chi^2_{n-p,1-\alpha_2}\right)$$

11.3 T statistics

Watch this video before beginning.

We can now develop T statistics. Consider the linear contrast $\hat{\boldsymbol{\beta}}^t\mathbf{t}$. First note that $\hat{\boldsymbol{\beta}}^t\mathbf{t}$ is $N(\boldsymbol{\beta}^t\mathbf{t}, \mathbf{t}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{t}\sigma^2)$. Furthermore, $\text{Cov}(\hat{\boldsymbol{\beta}}, \mathbf{e}) = \text{Cov}((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}, \{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\}\mathbf{Y}) = 0$ since $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\{\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\} = \mathbf{0}$. Thus the residuals and estimated coefficients are independent, implying that $\hat{\boldsymbol{\beta}}^t\mathbf{t}$ and S^2 are independent. Therefore,

$$\frac{\hat{\boldsymbol{\beta}}\mathbf{t} - \boldsymbol{\beta}\mathbf{t}}{\sqrt{\mathbf{t}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{t}\sigma^2}} \Big/ \sqrt{\frac{n-p}{\sigma^2}S^2/(n-p)} = \frac{\hat{\boldsymbol{\beta}}\mathbf{t} - \boldsymbol{\beta}\mathbf{t}}{\sqrt{\mathbf{t}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{t}S^2}}$$

is a standard normal divided by the square root of an independent Chi-squared over its degrees of freedom, thus is T distributed with $n - p$ degrees of freedom.

11.4 F tests

Watch this video before beginning.

Consider testing the hypothesis that $H_0 : \mathbf{K}\boldsymbol{\beta} = 0$ versus not equal for $\mathbf{K}$ of full row rank (say v). Notice that $\mathbf{K}\hat{\boldsymbol{\beta}} \sim N(\mathbf{K}\boldsymbol{\beta}, \mathbf{K}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}^t\sigma^2 0$ and thus

$$(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{K}\boldsymbol{\beta})^t\{\mathbf{K}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}^t\sigma^2\}^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{K}\boldsymbol{\beta})$$

is Chi-squared with v degrees of freedom. Furthermore, it is independent of $\mathbf{e}$ being a function of $\hat{\boldsymbol{\beta}}$. Thus:

$$(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{K}\boldsymbol{\beta})^t\{\mathbf{K}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}^t\}^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{K}\boldsymbol{\beta})/vS^2$$

forms the ratio of two independent Chi-squared random variables over their degrees of freedom, which is an F distribution.

11.5 Coding example

Watch this video before beginning.

Consider the `swiss` dataset. Let's first make sure that we can replicate the coefficient table obtained by R.

```
> ## First let's see the coeficient table
> fit = lm(Fertility ~ ., data = swiss)
> round(summary(fit)$coef, 3)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        66.915     10.706   6.250    0.000
Agriculture        -0.172      0.070  -2.448    0.019
Examination        -0.258      0.254  -1.016    0.315
Education          -0.871      0.183  -4.758    0.000
Catholic            0.104      0.035   2.953    0.005
Infant.Mortality    1.077      0.382   2.822    0.007
> # Now let's do it more manually
> x = cbind(1, as.matrix(swiss[,-1]))
> y = swiss$Fertility
> beta = solve(t(x) %*% x, t(x) %*% y)
> e = y - x %*% beta
> n = nrow(x); p = ncol(x)
> s = sqrt(sum(e^2) / (n - p))
> #Compare with lm
> c(s, summary(fit)$sigma)
[1] 7.165369 7.165369
> #calculate the t statistics
> betaVar = solve(t(x) %*% x) * s ^ 2
> ## Show that standard errors agree with lm
> cbind(summary(fit)$coef[,2], sqrt(diag(betaVar)))
                        [,1]        [,2]
(Intercept)      10.70603759 10.70603759
Agriculture       0.07030392  0.07030392
Examination       0.25387820  0.25387820
Education         0.18302860  0.18302860
Catholic          0.03525785  0.03525785
Infant.Mortality  0.38171965  0.38171965
> # Show that the tstats agree
> tstat = beta / sqrt(diag(betaVar))
> cbind(summary(fit)$coef[,3],  tstat)
                      [,1]      [,2]
(Intercept)       6.250229  6.250229
Agriculture      -2.448142 -2.448142
Examination      -1.016268 -1.016268
Education        -4.758492 -4.758492
Catholic          2.952969  2.952969
Infant.Mortality  2.821568  2.821568
> # Show that the P-values agree
> cbind(summary(fit)$coef[,4],  2 * pt(- abs(tstat), n - p)
                         [,1]         [,2]
(Intercept)      1.906051e-07 1.906051e-07
Agriculture      1.872715e-02 1.872715e-02
Examination      3.154617e-01 3.154617e-01
Education        2.430605e-05 2.430605e-05
Catholic         5.190079e-03 5.190079e-03
Infant.Mortality 7.335715e-03 7.335715e-03
> # Get the F statistic
> # Set K to grab everything except the intercept
> k = cbind(0, diag(rep(1, p - 1)))
> kvar = k %*% solve(t(x) %*% x) %*% t(k)
> fstat = t(k %*% beta) %*% solve(kvar) %*% (k %*% beta) / (p - 1) / s ^ 2
> #Show that it’s equal to what lm is giving
> cbind(summary(fit)$fstat, fstat)
> #Calculate the p-value
> pf(fstat, p - 1, n - p, lower.tail = FALSE)
             [,1]
[1,] 5.593799e-10
> summary(fit)
## ... only showing the one relevant line ...
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

11.6 Prediction intervals

It's worth discussing prediction intervals. The obvious prediction at a new set of covariates, $\mathbf{x}_0$, is $\mathbf{x}_0^t\hat{\boldsymbol{\beta}}$. This is then just a linear contrast of the $\boldsymbol{\beta}$ and so the interval would be

$$\mathbf{x}_0^t\hat{\boldsymbol{\beta}} \pm t_{n-p,1-\alpha/2}s\sqrt{\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}.$$

If you've taken an introductory regression class, you it will have noted the difference between a prediction interval and a confidence interval for a mean at a new value of $\mathbf{x}$. For a prediction interval, we want to estimate a range of possible values for y at that value of $\mathbf{x}$, a different statement than trying to estimate the average value of y at that value of $\mathbf{x}$. As we collect infinite data, we should get the average value exactly. However, predicting a new value involves intrinsic variability that can't go away no matter how much data we use to build our model.

As an example, imagine the difference between the following two tasks: guess the sale price of a diamond given its weight versus guess the average sale price of diamonds given a particular weight. With enough data, we should get the average sale price exactly. However, we still won't know exactly what the sale price of a diamond would be.

To account for this, we develop prediction intervals. These are not confidence intervals, because they are trying to estimate something random, not a fixed parameter. Consider estimating Y_0 at $\mathbf{x}$ value $\mathbf{x}_0$. Note that

$$\text{Var}(Y_0 - \mathbf{x}_0^t\hat{\boldsymbol{\beta}}) = (1 + \mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0)\sigma^2$$

For homework, show that

$$\frac{Y_0^t - \mathbf{x}_0^t\hat{\boldsymbol{\beta}}}{s\sqrt{1 + \mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0\sigma^2}}$$

follows a T distribution with $n-p$ degrees of freedom. Finish, by showing that

$$P\{y_0 \in [\mathbf{x}_0^t\hat{\boldsymbol{\beta}} \pm t_{n-p,1-\alpha/2}s\sqrt{1 + \mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}]\} = 1 - \alpha.$$

This is called a prediction interval. Notice the variability under consideration contains $\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0$, which goes to 0 as we get more $\mathbf{X}$ variability and 1, which represents the intrinsic part of the variability that doesn't go away as we collect more data.

11.6.1 Coding example

Let's try to predict a car's MPG from other characteristics.

```
> fit = lm(mpg ~ hp + wt, data = mtcars)
> newcar = data.frame(hp = 90, wt = 2.2)
> predict(fit, newdata = newcar)
       1
25.83648
> predict(fit, newdata = newcar, interval = "confidence")
```

```
       fit      lwr      upr
1 25.83648 24.46083 27.21212
> predict(fit, newdata = newcar, interval = "prediction")
       fit      lwr      upr
1 25.83648 20.35687 31.31609
> #Doing it manually
> library(dplyr)
> y = mtcars$mpg
> x = as.matrix(cbind(1, select(mtcars, hp, wt)))
> n = length(y)
> p = ncol(x)
> xtxinv = solve(t(x) %*% x)
> beta = xtxinv %*% t(x) %*% y
> x0 = c(1, 90, 2.2)
> yhat = x %*% beta
> e = y - yhat
> s = sqrt(sum(e^2 / (n - p)))
> yhat0 = sum(x0 * beta)
> # confidence interval
> yhat0 + qt(c(0.025, .975), n - p) * s * sqrt(t(x0) %*% xtxinv %*% x0)
[1] 24.46083 27.21212
> # prediction interval
> yhat0 + qt(c(0.025, .975), n - p) * s * sqrt(1 + t(x0) %*% xtxinv %*% x0)
[1] 20.35687 31.31609
```

11.7 Confidence ellipsoids

An hyper-ellipsoid with center $\mathbf{v}$ is defined as the solutions in $\mathbf{x}$ of $(\mathbf{x}-\mathbf{v})^t\mathbf{A}(\mathbf{x}-\mathbf{v}) = 1$. The eigenvalues of $\mathbf{A}$ determine the length of the axes of the ellipsoid. The set of points $\{\mathbf{x} \mid (\mathbf{x}-\mathbf{v})^t\mathbf{A}(\mathbf{x}-\mathbf{v}) \leq 1\}$ lie in the interior of the hyper-ellipsoid.

Now consider our F statistic from earlier on in the chapter:

$$(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})^t\{\mathbf{K}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}^t\}^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})/vS^2$$

We would fail to reject $H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{m}$ is less than the appropriate cut off from an F distribution, say $F_{1-\alpha}$. So, the set of points

$$\{\mathbf{m} \mid (\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})^t\{\mathbf{K}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}^t\}^{-1}(\mathbf{K}\hat{\boldsymbol{\beta}} - \mathbf{m})/vS^2F_{1-\alpha} \leq 1\}$$

forms a confidence set. From the discussion above, we see that this is a hyper ellipsoid. This multivariate form of confidence interval is called a confidence ellipse. These are of course most useful when the dimension is 2 or 3 so that we can visualize it as an actual ellipse.

```
fit = lm(mpg ~ disp + hp , mtcars)
```

```
open3d()
plot3d(ellipse3d(fit), col = "red", alpha = .5, aspect = TRUE)

## Doing it directly
beta = coef(fit)
Sigma = vcov(fit)
n = nrow(mtcars); p = length(beta)

A = Sigma * (3 * qf(.95, 3, n - p))
nms = names(beta)

open3d()
## Using the definition of an elipse
##(x - b)' A (x - b) = 1
plot3d(ellipse3d(A, centre = beta, t = 1),
                 color = "blue",  alpha = .5, aspect = TRUE,
       xlab = nms[1], ylab = nms[2], zlab = nms[3])

## Using the more statistical version
## Provide ellipse3d with the variance covariance matrix
plot3d(ellipse3d(Sigma, centre = beta, t = sqrt(3 * qf(.95, 3, n - p))),
                 color = "green",  alpha = .5, aspect = TRUE,
       xlab = nms[1], ylab = nms[2], zlab = nms[3], add = TRUE)
```

Chapter 12

Residuals revisited

12.1 Introduction to residuals

For a good treatment of residuals and the other topics in this chapter, see the book by Myers (Myers, 1990).

Now with some distributional results under our belt, we can discuss distributional properties of residuals. Note that, as a non-full rank linear transformation of normals, the residuals are singular normal. When $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, the mean of the residuals is $\mathbf{0}$, variance of the residuals is: given by

$$\text{Var}(\mathbf{e}) = \text{Var}\{(\mathbf{I} - \mathbf{H_X})\mathbf{y}\} = \sigma^2(\mathbf{I} - \mathbf{H_X}).$$

As a consequence, we see that the diagonal elements of $\mathbf{I} - \mathbf{H_X} \geq 0$ and thus the diagonal elements of $\mathbf{H_X}$ must be less than one. (A fact that we'll use later).

A problem with the residuals is that they have the units of $\mathbf{Y}$ and thus are not comparable across experiments. Taking

$$\text{Diag}\{S^2(I - \mathbf{H}_x)\}^{-1/2}\mathbf{e},$$

i.e., standardizing the residuals by their estimated standard deviation, does get rid of the units. However, the resulting quantities are not comparable to T-statistics since the numerator elements (the residuals) are not independent of S^2. The residuals standardized in this way are called “studentized” residuals. Studentized residuals are a standard part of most statistical software.

12.1.1 Coding example

```
> data(mtcars)
> y = mtcars$mpg
> x = cbind(1, mtcars$hp, mtcars$wt)
> n = nrow(x); p = ncol(x)
> hatmat =  x %*% solve(t(x) %*% x) %*% t(x)
> residmat =  diag(rep(1, n)) - hatmat
> e = residmat %*% y
> s = sqrt(sum(e^2) / (n - p))
> rstd = e / s / sqrt(diag(residmat))
> # compare with rstandard, r's function
> # for calculating standarized residuals
> cbind(rstd, rstandard(lm(y ~ x - 1)))
         [,1]        [,2]
1  -1.01458647 -1.01458647
2  -0.62332752 -0.62332752
3  -0.98475880 -0.98475880
4   0.05332850  0.05332850
5   0.14644776  0.14644776
6  -0.94769800 -0.94769800
...
```

12.2 Press residuals

Consider the model $\mathbf{y} \sim N(\mathbf{W}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$ where $\gamma = [\boldsymbol{\beta}^t \; \Delta_i]$, $\mathbf{W} = [\mathbf{X} \; \boldsymbol{\delta}_i]$ where $\boldsymbol{\delta}_i$ is a vector of all zeros except a 1 for row i. This model has a shift in position i, for example if there is an outlier at that position. The least squares criterion can be written as

$$\sum_{k \neq i}\left(y_k - \sum_{j=1}^{p} x_{kj}\beta_j\right)^2 + \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \Delta_i\right)^2. \tag{12.1}$$

Consider holding $\boldsymbol{\beta}$ fixed, then we get that the estimate of Δ_i must satisfy

$$\Delta_i = y_i - \sum_{j=1}^{p} x_{ij}\beta_j$$

and thus the right hand term of (12.1) is 0. Then we obtain $\boldsymbol{\beta}$ by minimizing

$$\sum_{k \neq i}\left(y_k - \sum_{j=1}^{p} x_{kj}\beta_j\right)^2.$$

Therefore $\hat{\boldsymbol{\beta}}$ is exactly the least squares estimate having deleted the i^{th} data point; notationally, $\hat{\boldsymbol{\beta}}^{(-i)}$. Thus, $\hat{\delta}_i$ is a form of residual obtained when deleting the i^{th} point from the fitting then comparing it to the fitted value,

$$\hat{\Delta}_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_k^{(-i)}.$$

Notice that the fitted value at the i^{th} data point is then $\sum_{j=1}^{p} x_{ij}\hat{\beta}_k^{(-i)} + \hat{\Delta}_i = \; y_i$ and thus the residual is zero. The term $\hat{\Delta}_i$ is called the PRESS residual, the difference between the observed value and the fitted value with that point deleted.

Since the residual at the i^{th} data point is zero, the estimated variance from thsi model is exactly equal to the variance estimate having removed the i^{th} data point. The t test for δ_i is then a form of standardized residual, that exactly follows a t distribution under the null hypothesis that $\delta_i = 0$.

12.2.1 Computing PRESS residuals

It is interesting to note that PRESS residuals don't actually require recalculating the model with the i^{th} datapoint deleted. Let $\mathbf{X}^t = [\mathbf{z}_1 \ \ldots \ \mathbf{z}_n]$ so that $\mathbf{z}_i$ is the i^{th} row of the matrix $\mathbf{z}$ (hence column i of $\mathbf{z}^t$). We use $\mathbf{z}$ for the rows, since we've already reserved $\mathbf{x}$ for the columns of $\mathbf{X}$. Notice, then that

$$\mathbf{X}^t\mathbf{X} = \sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^t.$$

Thus, $\mathbf{X}^{(-i),t}\mathbf{X}^{(-i)}$, the x transpose x matrix with the i^{th} data point deleted is simply

$$\mathbf{X}^{(-i),t}\mathbf{X}^{(-i)} = \mathbf{X}^t\mathbf{X} - \mathbf{z}_i\mathbf{z}_i^t.$$

We can appeal to the Sherman, Morrison, Woodbury theorem for the inverse (Wikipedia)

$$(\mathbf{X}^{(-i),t}\mathbf{X}^{(-i)})^{-1} = (\mathbf{X}^t\mathbf{X})^{-1} + \frac{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{z}_i\mathbf{z}_i^t(\mathbf{X}^t\mathbf{X})^{-1}}{1 - \mathbf{z}_i^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{z}_i}$$

Define h_{ii} as diagonal element i of $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ which is equal to $\mathbf{z}_i^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{z}_i$. (To see this, pre and post multiply this matrix by a vector of zeros with a one in the position i, an operation which grabs the i^{th} diagonal entry.) Furthermore, note that $\mathbf{X}^t\mathbf{y} = \sum_{i=1}^n \mathbf{z}_i y_i$ so that

$$\mathbf{X}^{(-i),t}\mathbf{y}^{(-i)} = \mathbf{X}^t\mathbf{y} - \mathbf{z}_i y_i.$$

Then we have that the predicted value for the i^{th} data point where it was not used in the fitting is:

$$\begin{aligned}
\hat{y}_i^{(-i)} &= \mathbf{z}_i^t(\mathbf{X}^{(-i),t}\mathbf{X}^{(-i)})^{-1}\mathbf{X}^{(-i),t}\mathbf{y}^{(-i)} \\
&= \mathbf{z}_i^t\left((\mathbf{X}^t\mathbf{X})^{-1} + \frac{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{z}_i\mathbf{z}_i^t(\mathbf{X}^t\mathbf{X})^{-1}}{1 - h_{ii}}\right)(\mathbf{X}^t\mathbf{y} - \mathbf{z}_i y_i) \\
&= \hat{y}_i + \frac{h_{ii}}{1-h_{ii}}\hat{y}_i - h_{ii}y_i - \frac{h_{ii}^2 y_i}{1-h_{ii}} \\
&= \frac{\hat{y}_i}{1-h_{ii}} + y_i - \frac{y_i}{1-h_{ii}}
\end{aligned}$$

So that we wind up with the equality:

$$y_i - \hat{y}_i^{(-i)} = \frac{y_i - \hat{y}_i}{1-h_{ii}} = \frac{e_i}{1-h_{ii}}$$

In other words, the PRESS residuals are exactly the ordinary residuals divided by $1 - h_{ii}$.

12.3 Externally studentized residuals

It’s often useful to have standardized residuals where a data point in question didn’t influence the residual variance. The normalized PRESS residuals are, as seen in 12.2. However, the PRESS residuals are leave one out residuals, and thus the i^{th} point was deleted for the fitted value. An alternative strategy is to normalize the ordinary residuals by dividing by a standard deviation estimate calculated with the i^{th} data point deleted. That is,

$$\frac{e_i}{s^{(-i)}\sqrt{1-h_{ii}}}.$$

In this statistic, observation i hasn’t had the opportunity to impact the variance estimate.

Given that the PRESS residuals are $\frac{e_i}{1-h_{ii}}$, their variance is $\sigma^2/\sqrt{1-h_{ii}}$. Then we have that the press residuals normalized (divided by their standard deviations) are

$$\frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

If we use the natural variance estimate for the press residuals, the estimated variance calculated with the i^{th} data point deleted, then the estimated normalized PRESS residuals are the same as the externally standardized residuals. As we know that these also arise out of the T-test for the mean shift outlier model from Section 12.2.

12.4 Coding example

First let’s use the `swiss` dataset to show how to calculate the ordinary residuals and show that they are the same as those output by `resid`.

```
> y = swiss$Fertility
> x = cbind(1, as.matrix(swiss[,-1]))
> n = nrow(x); p = ncol(x)
> hatmat = x %*% solve(t(x) %*% x) %*% t(x)
> ## ordinary residuals
> e = (diag(rep(1, n)) - hatmat) %*% y
> fit = lm(y ~ x)
> ## show that they’re equal by taking the max absolute difference
> max(abs(e - resid(fit)))
[1] 4.058975e-12
```

Next, we calculate the standardized residuals and show how to get them automatically with `rstandard`

```
> ## standardized residuals
> s = sqrt(sum(e ^ 2) / (n - p))
> rstd = e / s / sqrt(1 - diag(hatmat))
> ## show that they’re equal by taking the max absolute difference
> max(abs(rstd - rstandard(fit)))
[1] 6.638023e-13
```

Next, let's calculate the PRESS residuals both by leaving out the ith observation (in this case observation 6) and by the shortcut formula

```
> i = 6
> yi = y[i]
> yihat = predict(fit)[i]
> hii = diag(hatmat)[i]
> ## fitted model without the ith data point
> y.minus.i = y[-i]
> x.minus.i = x[-i,]
> beta.minus.i = solve(t(x.minus.i) %*% (x.minus.i)) %*% t(x.minus.i) %*% y.minus.i
> yhat.i.minus.i = sum(x[i,] * beta.minus.i)
> pressi = yi - yhat.i.minus.i
> c(pressi, e[i] / (1 - hii))
          Porrentruy
 -17.96269  -17.96269
```

Now show that the `rstudent` (externally studentized) residuals and normalized PRESS residuals are the same

```
> ## variance estimate with i deleted
> e.minus.i = y.minus.i - x.minus.i %*% beta.minus.i
> s.minus.i = sqrt(sum(e.minus.i ^ 2) / (n - p  - 1))
> ## show that the studentized residual is the PRESS residual standardized
> ei / s.minus.i / sqrt(1 - hii)
Porrentruy
 -2.367218
> rstudent(fit)[i]
        6
-2.367218
```

Finally, show that the mean shift outlier model residuals give the PRESS and the rstudent residuals.

```
> delta = rep(0, n); delta[i] = 1
> w = cbind(x, delta)
> round(summary(lm(y ~ w - 1))$coef, 3)
                  Estimate Std. Error t value Pr(>|t|)
w                   65.456     10.170   6.436    0.000
wAgriculture        -0.210      0.069  -3.067    0.004
wExamination        -0.323      0.242  -1.332    0.190
wEducation          -0.895      0.174  -5.149    0.000
wCatholic            0.113      0.034   3.351    0.002
wInfant.Mortality    1.316      0.376   3.502    0.001
wdelta             -17.963      7.588  -2.367    0.023
```

So notice that the the estimate for `wdelta` is the PRESS residual while the `t value` is the externally studentized residual.

Chapter 13

Under and overfitting

In linear models, we can characterize forms of under and overfitting. For this chapter consider the following:

Model 1: $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$

Model 2: $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$

where the $\boldsymbol{\epsilon}$ are assumed iid normals with variance σ^2. We further differentiate between the assumed model and the true model. If we assume Model 1 and Model 2 is true, we have underfit the model (omitted variables that were necessary). In contrast, if we assume Model 2 and Model 1 is true, we have overfit the model (included variables that were unnecessary).

13.1 Impact of underfitting

Consider underfitting the model. That is we errantly act as if Model 1 is true, but in fact model 2 is true. Such a situation would arise if there were unmeasured or unknown confounders. Then consider the bias of our estimate of $\beta 1$.

$$E[\hat{\boldsymbol{\beta}}_1] = E[(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{Y}] = (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) = \boldsymbol{\beta}_1 + (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{X}_2\boldsymbol{\beta}_2.$$

Thus, $(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{X}_2\boldsymbol{\beta}_2$ is the bias in estimating $\boldsymbol{\beta}_1$. Notice that there is no bias if $\mathbf{X}_1^t\mathbf{X}_2 = \mathbf{0}$. Consider the case where both design matrices are centered. Then $\frac{1}{n-1}\mathbf{X}_1^t\mathbf{X}_2$ is the empirical variance/covariance matrix between the columns of $\mathbf{X}_1$ and $\mathbf{X}_2$. Thus, if our omitted variables are uncorrelated with our included variables, then no bias exists. One way to try to force this in practice is to randomize the levels of the variables in $\mathbf{X}_1$. Then, the empirical correlation will be low with high probability. This is very commonly done when $\mathbf{X}_1$ contains only a single treatment indicator.

Our theoretical standard errors for the $\hat{\boldsymbol{\beta}}_1$ are still correct in that

$$\text{Var}(\hat{\boldsymbol{\beta}}_1) = (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\sigma^2.$$

However, we still have to estimate σ^2.

We can also see the impact of underfitting on the bias of residual variance estimation.

$$\begin{aligned}
E[(n-p_1)S^2] &= E[\mathbf{Y}^t(\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{Y}] \\
&= (\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2)^t\{\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\}(\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2) \\
&+ \mathrm{trace}[\{\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\}\sigma^2] \\
&= \boldsymbol{\beta}_2^t\mathbf{X}_2^t\{\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\}\mathbf{X}_2\boldsymbol{\beta}_2+(n-p_1)\sigma^2
\end{aligned}$$

Therefore S^2 is biased upward. It makes sense that we would tend to overestimate the residual variance if we've attributed to the error structure variation that is actually structured and due to unmodeled systematic variation.

13.2 Impact of overfitting

Consider now fitting Model 2 when, in fact, Model 1 is true. There is no bias in our estimate of $\boldsymbol{\beta}_1$, since we have fit the correct model; it's just $\boldsymbol{\beta}_2 = \mathbf{0}$.

13.3 Variance under an overfit model

If we fit Model 2, but Model 1 is correct, then our variance estimate is unbiased. We've fit the correct model, we just allowed the possibility that $\boldsymbol{\beta}_2$ was non-zero when it is exactly zero. Therefore S^2 is unbiased for σ^2. Recall too that

$$\frac{(n-p_1-p_2)S_2^2}{\sigma^2} \sim \chi^2_{n-p_1-p_2},$$

where the subscript 2 on S_2^2 is used to denote the fitting where Model 2 was asssumed. Similarly,

$$\frac{(n-p_1)S_1^2}{\sigma^2} \sim \chi^2_{n-p_1},$$

where S_1^2 is the variance assuming Model 1 is true. Using the fact that the variance of a Chi squared is twice the degrees of freedom, we get that

$$\frac{\mathrm{Var}(S_2^2)}{\mathrm{Var}(S_1^2)} = \frac{(n-p_1)^2}{(n-p_1-p_2)^2}.$$

Thus, despite both estimates being unbiased, the variance of the estimated variance under Model 2 is higher.

13.3.1 Variance inflation

Now consider Var$(\hat{\boldsymbol{\beta}}_1) = (\mathbf{X}^t\mathbf{X})^{-1}\sigma^2$, where $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$. Recall, that the estimate for $\hat{\boldsymbol{\beta}}_1$ can be obtained by regression of $\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2}$ on $\mathbf{e}_{\mathbf{y}|\mathbf{X}_2}$. Thus,

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{y}|\mathbf{X}_2}$$

Let $\mathbf{H}_{\mathbf{X}_2}$ be the hat matrix for $\mathbf{X}_2$. Thus,

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_1) &= (\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{I}-\mathbf{H}_{\mathbf{X}_2})\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\sigma^2 \\
&= (\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\sigma^2 - (\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{H}_{\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{x}_2}(\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\sigma^2 \\
&= (\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\sigma^2.
\end{aligned}$$

The latter term drops out since

$$\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{H}_{\mathbf{X}_2} = \mathbf{X}_1^t(\mathbf{I}-\mathbf{H}_{X_2})\mathbf{H}_{\mathbf{X}_2} = 0.$$

Consider any linear contrast $\mathbf{q}^t\boldsymbol{\beta}_1$ then

$$\begin{aligned}
\text{Var}(\mathbf{q}^t\hat{\boldsymbol{\beta}}_1) &= \mathbf{q}^t(\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\mathbf{q}\sigma^2 \\
&= \mathbf{q}^t(\mathbf{X}_1^t\mathbf{X}_1 - \mathbf{X}_1^t\mathbf{H}_{\mathbf{X}_2}\mathbf{X}_1)^{-1}\mathbf{q}\sigma^2 \\
&= \mathbf{q}^t\{(\mathbf{X}_1^t\mathbf{X}_1)^{-1} + \mathbf{W}\}\mathbf{q}\sigma^2
\end{aligned}$$

where $\mathbf{W}$ is a symmetric matrix. This latter identity can be obtained via the Woodbury theorem Wikipedia. Thus we can say that

$$\text{Var}(\mathbf{q}^t\hat{\boldsymbol{\beta}}_1) = \mathbf{q}^t\{(\mathbf{X}_1^t\mathbf{X}_1)^{-1} + \mathbf{W}\}\mathbf{q}\sigma^2 \geq \mathbf{q}^t(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{q}\sigma^2$$

Therefore, the variance assuming Model 2 will always be greater than the variance assuming Model 1. Note that at no point did we utilize which model was actually true. Thus we arrive at an essential point, adding more regressors into a linear model necessarily increases the standard error of the ones already included. This is called “variation inflation”. The estimated variances need not go up, since σ^2 will go down as we include variables. However, the central point is that one concern with including unnecessary regressors is inflating a component of the standard error needlessly.

Further note that σ^2 drops out in the ratio of the variances. We can thus exactly calculate the percentage increase in variance caused by including regressors. A particularly useful such summary is the variance inflation factor (VIF).

13.3.2 Variance inflation factors

Assume that $\mathbf{X}_1$ is a vector and that the intercept has been regressed out of both of $\mathbf{X}_1$ and $\mathbf{X}_2$. Recall from above that the variance for $\beta 1$ assuming Model 2 is (note β_1 is a scalar since we're assuming $\mathbf{X}_1$ is a vector)

$$\begin{aligned}
\text{Var}(\beta_1) &= (\mathbf{e}^t_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{e}_{\mathbf{X}_1|\mathbf{X}_2})^{-1}\sigma^2. \\
&= \frac{\sigma^2}{\mathbf{X}_1^t(\mathbf{I}-\mathbf{H}_{\mathbf{X}_2})\mathbf{X}_1} \\
&= \frac{\sigma^2}{\mathbf{X}_1^t\mathbf{X}_1^t} \times \frac{\mathbf{X}_1^t\mathbf{X}_1}{\mathbf{X}_1^t(\mathbf{I}-\mathbf{H}_{\mathbf{X}_2})\mathbf{X}_1}
\end{aligned}$$

Recall from partitioning sums of squares (remember that we've removed the intercept from both)

$$\mathbf{X}_1^t\mathbf{X}_1 = \mathbf{X}_1^t\mathbf{H}_{\mathbf{X}_2}\mathbf{X}_1 + \mathbf{X}_1^t(\mathbf{I}-\mathbf{H}_{\mathbf{X}_2})\mathbf{X}_1$$

and that $\frac{\mathbf{X}_1^t \mathbf{H}_{\mathbf{X}_2} \mathbf{X}_1}{\mathbf{X}_1^t \mathbf{X}_1}$ is the R^2 value for $\mathbf{X}_1$ as an outcome and $\mathbf{X}_2$ as a predictor. Let's call it R_1^2 so as not to confuse it with the R^2 calcualted with $\mathbf{Y}$ as an outcome. Then we can write

$$\text{Var}(\beta_1) = \frac{\sigma^2}{\mathbf{X}_1^t \mathbf{X}_1^t} \frac{1}{1 - R_1^2}.$$

Note that $R^2 = 1$ if $\mathbf{X}_2$ is orthogonal $\mathbf{X}_1$. Thus,

$$\frac{1}{1 - R_1^2}$$

Is the relative increase in variability in estimating β_1 comparing the data as it is to the ideal case where $\mathbf{X}_1$ is orthogonal to $\mathbf{X}_2$. Similarly, since $\frac{\sigma^2}{\mathbf{X}_1^t \mathbf{X}_1^t}$ is the variance if $\mathbf{X}_2$ is omitted from the model. So $1/(1 - R_1^2)$ is also the increase in the variance by adding the other regressors in $\mathbf{X}_2$.

This calculation can be performed for each regressor in turn. The $1/(1 - R^2)$ value for each regressor as an outcome with the remainder as predictors are the so-called Variance Inflation Factors (VIFs). They give information about how much addition variance is incurred by multicolinearity among the regressors.

13.3.3 Coding example

Let's look at variance inflation factors for the `swiss` dataset.

```
> library(car)
> data(swiss)
> fit4 = lm(Fertility ~ ., data = swiss)
> vif(fit4)
Agriculture      Examination        Education         Catholic Infant.Mortality
   2.284129         3.675420         2.774943         1.937160         1.107542
```

Thus, consider examination. The VIF of 3.7 suggest there's almost four times as much variability in estimating the Examination coefficient by the inclusion of the other variables. We can show the calculation of these statistics manually as such.

```
1 / (1 - summary(lm(Examination ~ . - Fertility, data = swiss))$r.squared)
[1] 3.67542
```

Consider comparing the estimated standard errors for the examination variable

```
> summary(lm(Fertility ~ Examination, data = swiss))$coef
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 86.818529  3.2576034 26.651043 3.353924e-29
Examination -1.011317  0.1781971 -5.675275 9.450437e-07
> summary(lm(Fertility ~ ., data = swiss))$coef
                   Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)      66.9151817 10.70603759  6.250229 1.906051e-07
```

```
Agriculture      -0.1721140  0.07030392 -2.448142 1.872715e-02
Examination      -0.2580082  0.25387820 -1.016268 3.154617e-01
Education        -0.8709401  0.18302860 -4.758492 2.430605e-05
Catholic          0.1041153  0.03525785  2.952969 5.190079e-03
Infant.Mortality  1.0770481  0.38171965  2.821568 7.335715e-03
```

Here the increase in variance is `(0.25387820 / 0.1781971)` squared which is approximately 2. This is much less than is predicted by the VIF because it involves the estimated variance rather than the actual variance.

Chapter 14

Parameter estimability and penalties

In this section we consider parameter estimability and penalties.

14.1 Estimability

This section draws heavily from the wonderful book by Searle (2012).

We define a linear combination of the slope parameters, $\mathbf{q}^t\boldsymbol{\beta}$, as being estimable if it is equal to a linear combination of the expected value of $\mathbf{Y}$. In other words, $\mathbf{q}^t\boldsymbol{\beta}$ is estimable if it is equal to $\mathbf{t}^t E[\mathbf{Y}]$ for some value of $\mathbf{t}$.

I find estimability most useful when $\mathbf{X}$ is over-specified (not full rank). For example, consider an ANOVA model

$$Y_{ij} = \mu + \beta_i + \epsilon_{ij}.$$

Verify for yourself that the $\mathbf{X}$ matrix from this model is not full rank.

Because $\mathbf{t}^t E[\mathbf{Y}] = \mathbf{t}^t \mathbf{X}\boldsymbol{\beta}$ for all possible $\boldsymbol{\beta}$, $\mathbf{q} = \mathbf{t}^t\mathbf{X}$ and we obtain that estimable contrasts are necessarily linear combinations of the rows of the design matrix.

The most useful result in estimability is the invariance properties of estimable contrasts. Consider an not full rank design matrix. Then any solution to the normal equations:

$$\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{Y}$$

will minimize the least squares criteria (or equivalently maximize the likelihood under spherical Gaussian assumptions). (If you don't see this, verify it yourself using the tools from the first few chapters.) Since $\mathbf{X}$ is not full rank, this will have infinite solutions. Let $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ be any two such solutions. For estimable quantities, $q^t\hat{\boldsymbol{\beta}} = q^t\tilde{\boldsymbol{\beta}}$. That is, the particular solution to the normal doesn't matter for estimable quantities. This should be clear given the definition of estimability. Recall that least squares projects onto the plane defined by linear combinations of the columns of $\mathbf{X}$. The projection, $\hat{\mathbf{Y}}$, is unique, while the particular linear combination is not in this case.

To discuss further. Suppose $\mathbf{q}^t\hat{\boldsymbol{\beta}} \neq \mathbf{q}^t\tilde{\boldsymbol{\beta}}$ for two solutions to the normal equations, $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ and estimable $\mathbf{q}^t\boldsymbol{\beta}$. Then $\mathbf{t}^t\mathbf{X}\hat{\boldsymbol{\beta}} \neq \mathbf{t}^t\mathbf{X}\tilde{\boldsymbol{\beta}}$. Let $\hat{\mathbf{Y}}$ be the projection of $\mathbf{Y}$ on the space of linear combinations of the columns of $\mathbf{X}$. However, since both are projections, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\tilde{\boldsymbol{\beta}}$. Multiplying by $\mathbf{t}^t$ then yields a contradiction.

14.1.1 Why it's useful

Consider the one way ANOVA setting again.

$$Y_{ij} = \mu + \beta_i + \epsilon_{ij}.$$

For $i = 1, 2$, $j = 1, \ldots, J$. One can obtain parameter identifiability by setting $\beta_2 = 0$, $\beta_1 = 0$, $\mu = 0$ or $\beta_1 + \beta_2 = 0$ (or one of infinitely many other linear contrasts). These constraints don't change the column space of the $\mathbf{X}$ matrix. (Thus the projection stays the same.) Recall that $\hat{y}_{ij} = \bar{y}_i$. Estimable functions are linear combinations of $E[Y_{ij}] = \mu + \beta_i$. So, note that

$$E[Y_{21}] - E[Y_{11}] = \beta_2 - \beta_1$$

is estimable and it will always be estimated by $\bar{y}_2 - \bar{y}_1$. Thus, regardless of which linear constraints one points on the model to achieve identifiability, the difference in the means will have the same estimate.

This also gives us a way to go between estimates with different constraints without refitting the models. Since for two sets of constraints we have:

$$\bar{y}_i = \hat{\mu} + \hat{\beta}_i = \tilde{\mu} + \tilde{\beta}_i,$$

yielding a simple system of equations to convert between estimates with different constraints.

14.2 Linear constraints

Consider performing least squares under the full row rank linear constraints

$$\mathbf{K}^t \boldsymbol{\beta} = \mathbf{z}.$$

One could obtain these estimates using Lagrange multipliers

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + 2\boldsymbol{\lambda}^t(\mathbf{k}^t\boldsymbol{\beta} - \mathbf{Z}) = \mathbf{y}^t\mathbf{y} - 2\boldsymbol{\beta}^2\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\lambda}^t(\mathbf{K}^t\boldsymbol{\beta} - \mathbf{z}).$$

Taking a derivative with respect to lambda yields

$$2(\mathbf{K}^t\boldsymbol{\beta} - \mathbf{z}) = 0 \tag{14.1}$$

Taking a derivative with respect to $\boldsymbol{\beta}$ we have:

$$-2\mathbf{X}^t\mathbf{y} + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + 2\mathbf{K}\boldsymbol{\lambda} = 0$$

which has a solution in $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{y} - \mathbf{K}\boldsymbol{\lambda}) = \hat{\boldsymbol{\beta}} - (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}\boldsymbol{\lambda}, \tag{14.2}$$

where $\hat{\boldsymbol{\beta}}$ is the OLS (unconstrained) estimate. Multiplying by $\mathbf{K}^t$ and using (14.1) we have that

$$\mathbf{z} = \mathbf{K}^t\hat{\boldsymbol{\beta}} - \mathbf{K}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}\boldsymbol{\lambda}$$

yielding a solution for $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda} = \{\mathbf{K}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}\}^{-1}(\mathbf{K}^t\hat{\boldsymbol{\beta}} - \mathbf{z}).$$

Plugging this back into (14.2) yields the solution:

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}\{\mathbf{K}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}\}^{-1}(\mathbf{K}^t\hat{\boldsymbol{\beta}} - \mathbf{z}).$$

Thus, one can fit constrained least squares estimates without actually refitting the model. Notice, in particular, that if one where to multiply this estimate by $\mathbf{K}^t$, the result would be $\mathbf{z}$.

14.2.1 Likelihood ratio tests

One can use this result to derive likelihood ratio tests of $H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{z}$ versus the general alternative. From the previous section, under the null hypothesis, the estimate under the null hypothesis,

$$\hat{\boldsymbol{\beta}}_{H_0} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}\{\mathbf{K}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}\}^{-1}(\mathbf{K}^t\hat{\boldsymbol{\beta}} - \mathbf{z}).$$

Of course, under the alternative, the estimate is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$. In both cases, the maximum likelihood variance estimate is $\frac{1}{n}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ with $\boldsymbol{\beta}$ as the estimate under either the null or alternative hypothesis. Let $\hat{\sigma}^2_{H_0}$ and $\hat{\sigma}^2$ be the two estimates.

The likelihood ratio statistic is

$$\frac{\mathcal{L}(\hat{\boldsymbol{\beta}}_{H_0}, \hat{\sigma}^2_{H_0})}{\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} = \left(\frac{\hat{\sigma}^2_{H_0}}{\hat{\sigma}^2}\right)^{-n/2}.$$

This is monotonically equivalent to $n\hat{\sigma}^2/n\hat{\sigma}^2_{H_0}$. However, we reject if the null is less supported than the alternative, i.e. this statistic is small, so we could equivalently reject if $n\hat{\sigma}^2_{H_0}/n\hat{\sigma}^2$ is large. Further note that

$$\begin{aligned}
n\hat{\sigma}^2_{H_0} &= ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{H_0}||^2 \\
&= ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{H_0}||^2 \\
&= ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 + ||\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{H_0}||^2 \\
&= n\hat{\sigma}^2 + ||\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{H_0}||^2
\end{aligned}$$

Notationally, let

$$SS_{reg} = ||\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{H_0}||^2 = ||\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{H_0}||^2$$

and $SS_{res} = n\hat{\sigma}^2$. The note that the inverse of our likelihood ratio is monotonically equivalent to $\frac{SS_{reg}}{SS_{res}}$

However, SS_{reg}/σ^2 and SS_{res}/σ^2 are both independent Chi-squared random variables with degrees of freedom $Rank(\mathbf{K})$ and $n - p$ under the null. (Prove this for homework.) Thus, our likelihood ratio statistic can exactly be converted into the F statistic of section 11.4. We leave the demonstration that the two are identical as a homework exercise.

This line of thinking can be extended. Consider the sequence of hypotheses:

$$\begin{aligned}
&H_1 : \mathbf{K}_1\boldsymbol{\beta} = \mathbf{z}\\
&H_2 : \mathbf{K}_2\mathbf{K}_1\boldsymbol{\beta} = \mathbf{K}_2\mathbf{z}\\
&H_3 : \mathbf{K}_3\mathbf{K}_2\mathbf{K}_1\boldsymbol{\beta} = \mathbf{K}_3\mathbf{K}_2\mathbf{z}\\
&\vdots
\end{aligned}$$

Each $\mathbf{K}_i$ is assumed full row rank and of fewer rows than $\mathbf{K}_{i-1}$. These hypotheses are nested with H_1 being the most restrictive, H_2 being the second most, and so on. (Note, if H_1 holds then H_2 holds but not vice versa.) Consider testing H_1 (null) versus H_2 (alternative). Note that under our general specification, discussing this problem will apply to testing H_i versus H_j. Under the arguments above, our likelihood ratio statistic will work out to be inversely equivalent to the statistic: $n\hat{\sigma}^2_{H_1}/n\hat{\sigma}^2_{H_2}$.

Further note that

$$\begin{aligned}
n\hat{\sigma}^2_{H_1} &= ||\mathbf{Y} - \hat{\mathbf{Y}}_{H_1}||\\
&= ||\mathbf{Y} - \hat{\mathbf{Y}}_{H_2}||^2 + ||\hat{\mathbf{Y}}_{H_2} - \hat{\mathbf{Y}}_{H_1}||^2 + 2(\mathbf{Y} - \hat{\mathbf{Y}}_{H_2})^t(\hat{\mathbf{Y}}_{H_2} - \hat{\mathbf{Y}}_{H_1})\\
&= ||\mathbf{Y} - \hat{\mathbf{Y}}_{H_2}||^2 + ||\hat{\mathbf{Y}}_{H_2} - \hat{\mathbf{Y}}_{H_1}||^2\\
&= SS_{RES}(H_2) + SS_{REG}(H_1 \mid H_2)
\end{aligned}$$

Here the cross product term in the second line is zero by (tedious yet straightforward) algebra and the facts that: $\mathbf{K}_2\mathbf{K}_1\hat{\boldsymbol{\beta}}_{H_1} = \mathbf{K}_2\mathbf{K}_1\hat{\boldsymbol{\beta}}_{H_2} = \mathbf{K}_2\mathbf{z}$ and $\mathbf{e}^t\mathbf{X} = \mathbf{0}$.

Thus, our likelihood ratio statistic is monotically equivalent to

$$SS_{REG}(H_1 \mid H_2)/SS_{RES}(H2).$$

Furthermore, Using the developed methods in the class the numerator is Chi-Squared with $Rank(\mathbf{K}_1)$ degrees of freedom, while the denominator has $n - \{Rank(\mathbf{K}_1) - Rank(\mathbf{K}_2)\}$ degrees of freedom, and they are independent. Thus we can construct an F test for nested linear hypotheses.

This process can be iterated, decomposing $SS_{RES}(H_2)$, so that:

$$n\hat{\sigma}^2_{H_1} = SS_{REG}(H_1 \mid H_2) + SS_{REG}(H_2 \mid H_3) + SS_{RES}(H_3)$$

And it could be iterated again so that:

$$n\hat{\sigma}^2_{H_1} = SS_{REG}(H_1 \mid H_2) + SS_{REG}(H_2 \mid H_3) + \ldots SS_{RES}(H_p)$$

where $SS_{RES}(H_p)$ is the residual sums of squares under the most elaborate model considered. The sums of squares add so that, for example,

$$SS_{REG}(H_1 \mid H_3) = SS_{REG}(H_1 \mid H_2) + SS_{REG}(H_2 \mid H_3)$$

and

$$SS_{RES}(H_3) = SS_{REG}(H_3 \mid H_4) + \ldots + SS_{RES}(H_4).$$

Thus, one could test any subset of the nested hypotheses by appropriately adding the sums of squares.

14.2.2 Example use

The most popular use of the general linear hypothesis is to consider nested hypotheses. That is, consider a linear model where $\boldsymbol{\beta}^t = [\beta_0 \; \beta_1 \; \ldots \; \beta_p]$ so that the β_i are ordered in decreasing scientific importance.

$$
\begin{aligned}
&H_1 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\
&H_2 : \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\
&H_3 : \beta_3 = \ldots = \beta_p = 0 \\
&\vdots \\
&H_p : \beta_p = 0
\end{aligned}
$$

Then testing H_1 versus H_2 tests whether β_1 is zero under the assumption that all of the remaining coefficients (excepting the intercept) are zero. Testing H_2 versus H_5 tests whether $\beta_2 = \beta_3 = \beta_4 = 0$ under the assumption that β_5 through β_p are 0.

14.2.3 Coding examples

Let's go through an example of fitting multiple models. We'll look at the swiss dataset. The following code fits three models for the dataset. First, we model the outcome, regional fertility, as a function of various aspects of the region. Imagine if we are particularly interested in agriculture as a variable. We fit three models: one a linear regression with just agriculture, then one including educational level variables (examination and education) and then one including all of the previous variables plus information on religion (percent Catholic) and infant mortality rates.

```
data(swiss)
fit1 = lm(Fertility ~ Agriculture, data = swiss)
fit2 = update(fit1, Fertility ~ Agriculture + Examination + Education)
fit3 = update(fit1, Fertility ~ Agriculture + Examination + Education + Catholic + Infan
anova(fit1, fit2, fit3)
```

The anova comand gets the relevant sums of squares for each of the models, resulting in the output

```
Analysis of Variance Table

Model 1: Fertility ~ Agriculture
Model 2: Fertility ~ Agriculture + Examination + Education
Model 3: Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     45 6283.1
2     43 3180.9  2    3102.2 30.211 8.638e-09 ***
3     41 2105.0  2    1075.9 10.477 0.0002111 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here, it would appear that inclusion of the other variables is necessary. However, let's see if we can create these sums of squares manually using our approach.

```
> xtilde = as.matrix(swiss);
> y = xtilde[,1]
> x1 = cbind(1, xtilde[,2])
> x2 = cbind(1, xtilde[,2:4])
> x3 = cbind(1, xtilde[,-1])
> makeH = function(x) x %*% solve(t(x) %*% x) %*% t(x)
> n = length(y); I = diag(n)
> h1 = makeH(x1)
> h2 = makeH(x2)
> h3 = makeH(x3)
> ssres1 = t(y) %*% (I - h1) %*% y
> ssres2 = t(y) %*% (I - h2) %*% y
> ssres3 = t(y) %*% (I - h3) %*% y
> ssreg2g1 = t(y) %*% (h2 - h1) %*% y
>ssreg3g2 = t(y) %*% (h3 - h2) %*% y
> out = rbind( c(n - ncol(x1), ssres1,                NA,      NA),
               c(n - ncol(x2), ssres2, ncol(x2) - ncol(x1), ssreg2g1),
               c(n - ncol(x3), ssres3, ncol(x3) - ncol(x2), ssreg3g2)
  )
> out
     [,1]     [,2] [,3]     [,4]
[1,]   45 6283.116   NA       NA
[2,]   43 3180.925    2 3102.191
[3,]   41 2105.043    2 1075.882
```

It is interesting to note that the F test comapring Model 1 to Model 2 from the `anova` command is obtained by dividing `3102.191 / 2` (a chi-squared divided by its 2 degrees of freedom) by `2105.043 / 41` (an independent chi-squared divided by its 3 degrees of freedom). The denominator of the F statistic is then the residual sum of squares from Model 3, not from Model 2.

This is why the following give two different answers for the F statistic:

```
> anova(fit1, fit2)
Analysis of Variance Table

Model 1: Fertility ~ Agriculture
Model 2: Fertility ~ Agriculture + Examination + Education
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     45 6283.1
```

```
2     43 3180.9  2    3102.2 20.968 4.407e-07 ***
---
> anova(fit1, fit2, fit3)
Analysis of Variance Table

Model 1: Fertility ~ Agriculture
Model 2: Fertility ~ Agriculture + Examination + Education
Model 3: Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     45 6283.1
2     43 3180.9  2    3102.2 30.211 8.638e-09 ***
3     41 2105.0  2    1075.9 10.477 0.0002111 ***
```

In the first case, the denominator of the F statistic is `3180.9 / 43`, the residual mean squared error for Model 2, as opposed to the latter case where it is dividing by the residual mean squared error for Model 3. Of course, under the null hypothesis, either approach yields an independent chi squared statistic in the denominator. However, using the Model 3 residual mean squared error reduces the denominator degrees of freedom, though also necessarily reduces the residual sum of squared errors (since extra terms in the regression model always do that).

14.3 Ridge regression

Consider quadratic constraints to least squares.

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \boldsymbol{\beta}^t\boldsymbol{\Gamma}\boldsymbol{\beta}.$$

In this case we consider instances where $\mathbf{X}$ is not necessarily full rank. The addition of the penalty is called “Tikhonov regularization” for the mathematician of that name. The specific instance of this regularization in regression is called ridge regression. The matrix $\boldsymbol{\Gamma}$ is typically assumed known or set to $\gamma\mathbf{I}$.

Another way to envision ridge regression is to think in the terms of a posterior mode on a regression model. Specifically, $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}/\sigma^2$ and consider the model where $\mathbf{y} \mid \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Then one obtains the posterior for $\boldsymbol{\beta}$ and σ by multiplying the two densities. The posterior mode would be obtained by minimizing minus twice the log of this product

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2/\sigma^2 + \boldsymbol{\beta}^t\boldsymbol{\Gamma}\boldsymbol{\beta}/\sigma^2.$$

which is equivalent to above in the terms of maximization for $\boldsymbol{\beta}$.

We'll leave it as an exercise to obtain that the estimate actually obtained is

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^t\mathbf{X} + \boldsymbol{\Gamma})^{-1}\mathbf{X}^t\mathbf{Y}.$$

To see how this regularization helps with invertibility of $\mathbf{X}^t\mathbf{X}$, consider the case where $\boldsymbol{\Gamma} = \gamma\mathbf{I}$. If γ is very large then $\mathbf{X}^t\mathbf{X} + \gamma\mathbf{I}$ is simply small numbers added around an identity matrix, which is clearly invertible.

Consider the case where $\mathbf{X}$ is column centered and is of full column rank. Let $\mathbf{U}\mathbf{D}\mathbf{V}^t$ be the SVD of $\mathbf{X}$ where $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}$. Note then $\mathbf{X}^t\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^t$ and $(\mathbf{X}^t\mathbf{X})^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^t$ so that the ordinary least squares estimate satisfies

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^t\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^t\mathbf{V}\mathbf{D}\mathbf{U}\mathbf{Y} = \mathbf{U}\mathbf{U}^t\mathbf{Y}.$$

Consider now the fitted values under ridge regression with $\mathbf{\Gamma} = \gamma\mathbf{I}$:

$$\begin{aligned}
\hat{\mathbf{Y}}_{ridge} &= \mathbf{X}(\mathbf{X}^t\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^t(\mathbf{V}\mathbf{D}^2\mathbf{V}^t + \gamma\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^t(\mathbf{V}\mathbf{D}^2\mathbf{V}^t + \gamma\mathbf{V}\mathbf{V}^t)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^t\{\mathbf{V}(\mathbf{D}^2 + \gamma\mathbf{I})\mathbf{V}^t\}^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^t\mathbf{V}(\mathbf{D}^2 + \gamma\mathbf{I})^{-1}\mathbf{V}^t\mathbf{V}\mathbf{D}\mathbf{U}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \gamma\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{W}\mathbf{U}^t\mathbf{Y}
\end{aligned}$$

where the third line follows since $\mathbf{X}$ is full column rank so that $\mathbf{V}$ is $p \times p$ of full rank and $\mathbf{V}^{-1} = \mathbf{V}^t$ so that $\mathbf{V}^t\mathbf{V} = \mathbf{V}\mathbf{V}^t = \mathbf{I}$. Here $\mathbf{W}$ is a diagonal matrix with elements

$$\frac{D_i^2}{D_i^2 + \gamma}$$

where D_i^2 are the eigenvalues.

In the not full rank case, the same identity can be found, though it takes a bit more work. Now assume that $\mathbf{X}$ is of full row rank (i.e. that $n < p$ and there are no redundant subjects). Now note that $\mathbf{V}$ does not have an inverse, while $\mathbf{U}$ does (and $\mathbf{U}^{-1} = \mathbf{U}^t$. Further note via the Woodbury theorem (where $\theta = 1/\lambda)d$:

$$\begin{aligned}
(\mathbf{X}^t\mathbf{X} + \gamma\mathbf{I})^{-1} &= \theta\mathbf{I} - \theta^2\mathbf{X}^t(\mathbf{I} + \theta\mathbf{X}\mathbf{X}^t)^{-1}\mathbf{X} \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\mathbf{D}\mathbf{U}^t(\mathbf{U}\mathbf{U}^t + \theta\mathbf{U}\mathbf{D}^2\mathbf{U}^t)^{-1}\mathbf{U}\mathbf{D}\mathbf{V}^t \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\mathbf{D}\mathbf{U}^t\{\mathbf{U}(\mathbf{I} + \theta\mathbf{D}^2)\mathbf{U}^t)\}^{-1}\mathbf{U}\mathbf{D}\mathbf{V}^t \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\mathbf{D}\mathbf{U}^t\{\mathbf{U}(\mathbf{I} + \theta\mathbf{D}^2)^{-1}\mathbf{U}^t)\}\mathbf{U}\mathbf{D}\mathbf{V}^t \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\mathbf{D}(\mathbf{I} + \theta\mathbf{D}^2)^{-1}\mathbf{D}\mathbf{V}^t \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\tilde{\mathbf{D}}\mathbf{V}^t
\end{aligned}$$

where $\tilde{\mathbf{D}}$ is diagonal with entries $D_i^2/(1 + \theta D_i^2)$ where D_i are the diagonal entries of $\mathbf{D}$. Then:

$$\begin{aligned}
\hat{\mathbf{Y}}_{Ridge} &= \mathbf{X}(\mathbf{X}^t\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^t(\theta\mathbf{I} - \theta^2\mathbf{V}\tilde{\mathbf{D}}\mathbf{V}^t)\mathbf{V}\mathbf{D}\mathbf{U}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}(\theta\mathbf{I} - \theta^2\tilde{\mathbf{D}})\mathbf{D}\mathbf{U}^t\mathbf{Y} \\
&= \mathbf{U}\mathbf{W}\mathbf{U}^t\mathbf{Y}
\end{aligned}$$

Thus we've covered the full row and column rank cases. (Omitting the instance where $\mathbf{X}$ is neither full row nor column rank.)

14.3.1 Coding example

In the example below, we use the `swiss` data set to illustrate fitting ridge regression. In this example, penalization isn’t really necessary, so the code is more used to simply show the fitting. Notice that `lm.ridge` and our code give slightly different answers. This is due to different scaling options for the design matrix.

```
data(swiss)
y = swiss[,1]
x = swiss[,-1]
y = y - mean(y)
x = apply(x, 2, function(z) (z - mean(z)) / sd(z))
n = length(y); p = ncol(x)
##get ridge regression estimates for varying lambda
lambdaSeq = seq(0, 100, by = .1)
betaSeq = sapply(lambdaSeq, function(l) solve(t(x) %*% x + l * diag(rep(1, p)), t(x) %*%
plot(range(lambdaSeq), range(betaSeq), type = "n", xlab = "- lambda", ylab = "Beta")
for (i in 1 : p) lines(lambdaSeq, betaSeq[i,])

##Use R’s function for Ridge regression
library(MASS)
fit = lm.ridge(y ˜ x, lambda = lambdaSeq)
plot(fit)
```

14.4 Lasso regression

The Lasso has been somewhat of a revolution of sorts in statistics and biostatistics of late. The central idea of the lasso is to create a penalty that forces coefficients to be zero. For centered $\mathbf{Y}$ and centered and scaled $\mathbf{X}$, consider minimizing

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$$

subject to $\sum_{i=1}^{p} |\beta_i| < t$. The Lagrangian form of this minimization can be written as minimizing

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \sum_{i=1}^{n} |\beta_i|.$$

Here $\boldsymbol{\lambda}$ is a penalty parameter. As the Lasso constrains $\sum_{i=1}^{p} |\beta_i| < t$, which has sharp corners on the axes, it has a tendency to set parameters exactly to zero. Thus, it is thought of as doing model selection along with penalization. Moreover, the Lasso handles the $p > n$ problem. Finally, it’s a convex optimization problem, so that numerically solving for the Lasso is stable. We can more generally specify the parameter as

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \sum_{i=1}^{n} |\beta_i|^q.$$

for $q > 0$. We obtain a case of ridge regression when $q = 2$ and the Lasso when $q = 1$. Since $(\sum_{i=1}^{n} |\beta_i|^q)^{1/q}$ is a norm, usually called the l_q norm, the various forms of regression are often calld l_q regression. For example, ridge regression could be called l_2 regression, the Lasso $\mathcal{L}_1$ regression and so on. We could write the penalized regression estimate as

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_q^q$$

where $||\cdot||_q$ is the l_q norm.

You can visualize the parameters easily using Wolfram's alpha: $|x_1| + |x_2|$, $|x_1|^2 + |x_2|^2$, $|x_1|^0.5 + |x_2|^0.5$, and $|x_1|^4 + |x_2|^4$. Notice that as q tends to zero, it tends to all of the mass on the axes where as q tends to infinity, it tends to a square. The limit as q tends to 0 is called the l_0 norm, which just penalizes the number of non-zero coefficients.

Just like with ridge regression, the Lasso has a Bayesian representation. Let the prior on β_i be iid from a Laplacian distribution with mean 0, which has density $\frac{\theta}{2}\exp(-\theta|\beta_i|)$, and is denoted Lapplace$(0, \theta)$. Then, the Lasso estimate is the posterior mean assuming $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$ and $\beta_i \sim_{iid}$ Laplace$(0, \lambda/2\sigma^2)$. Then minus twice the log of the posterior for $\boldsymbol{\beta}$, conditioning on σ, is proportional to

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_1.$$

The connection with Bayesian statistics is somewhat loose for lasso regression. While the Lasso is the posterior mode under a specific prior, whether or not that prior makes sense from a Bayesian perspective is not clear. Furthermore, the full posterior for a parameter in the model is averaged over several sparse models, so is actually not sparse. Also, the posterior mode is conditioned on σ under these assumptions, Bayesian analysis usually take into account the full posterior.

14.4.1 Coding example

Let's give an example of coding the Lasso. Here, because the optimization problem isn't closed form, we'll rely on the `lars` package from Tibshirani and Efron. Also assume the code from the ridge regression exmaple.

```
library(lars)
fit2 = lars(x, y, type = c("lasso"))
plot(fit2)
```

Chapter 15

Asymptotics

15.1 Convergence in probability

A series of random variables, Y_n, is said the converge in probability to a constant c if $P(|Y_n - c| > \epsilon) \rightarrow 0$ for any ϵ. A standard result is that convergence in probability to the mean is implied if the variance of random variable goes to zero (a consequence of Chebyshev's inequality). Specifically, let $Z_n = Y_n - \mu$ have mean 0, variance σ_n^2 and distribution F_n

$$\begin{aligned}
P(|Z_n| \geq \epsilon) &= \int_{|z_n| \geq \epsilon} dF_n(z_n) \\
&= \int_{z_n^2/\epsilon^2 \geq 1} dF_n(z_n) \\
&\leq \int_{z_n^2/\epsilon^2 \geq 1} \frac{z_n^2}{\epsilon^2} dF_n(z_n) \\
&\leq \int \frac{z_n^2}{\epsilon^2} dF_n(z_n) \\
&= \sigma_n^2/\epsilon^2.
\end{aligned}$$

Thus, according to our definition, Z_n converges in probability to 0 (thus Y_n converges in probability to μ) if the sequence of variances tends to zero.

Consider now convergence in probability of our slope estimate

$$\hat{\boldsymbol{\beta}}_n = (\mathbf{X}_n^t \mathbf{X}_n)^{-1} \mathbf{X}_n^t \mathbf{Y}_n$$

where subscripts have been added to denote the dependence on the sample size. This estimator is unbiased for all n. Under the assumption of iid errors, with a finite variance of $\mathbf{Y}_n$ of $\mathbf{I}\sigma^2$, the variance of a linear contrast of $\hat{\boldsymbol{\beta}}_n$ is

$$\mathbf{q}^t (\mathbf{X}_n^t \mathbf{X}_n)^{-1} \mathbf{q} \sigma^2.$$

Thus a sufficient condition for consistency of $\mathbf{q}^t \hat{\boldsymbol{\beta}}_n$ is for $\mathbf{q}^t (\mathbf{X}_n^t \mathbf{X}_n)^{-1} \mathbf{q}$ to converge to zero. Probably more useful is if sample variance covariance associated with the $\mathbf{X}_n$ converges, then the estimate is consistent for all linear contrasts.

In the particular case for linear regression, recall that the variance of the slope is

$$\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s^2_{x,n}}$$

where $s^2_{x,n}$ is the sample variance of the x's. Thus, as long as $s^2_{x,n}$ converges, the estimate is consistent. Alternatively, if the X_i are bounded, then the estimate will converge.

Consider now the case where $E[\mathbf{Y}_n] = \mathbf{X}_n\boldsymbol{\beta}$ but Var$(\mathbf{Y}_n) = \boldsymbol{\Sigma}_n$. Consider a working covariance matrix, $\mathbf{W}_n$, and the estimate

$$\hat{\boldsymbol{\beta}}(\mathbf{W}_n) = (\mathbf{X}_n^t \mathbf{W}_n^{-1} \mathbf{X}_n)^{-1} \mathbf{X}_n^t \mathbf{W}_n^{-1} \mathbf{Y}_n.$$

The OLS estimate is the case where $\mathbf{W}_n = \mathbf{I}$. (Notice that the estimate is invariant to scale changes in $\mathbf{W}_n$.). Notice that $\hat{\boldsymbol{\beta}}(\mathbf{W}_n)$ is unbiased for all n regardless of $\mathbf{W}_n$. The variance of $\hat{\boldsymbol{\beta}}(\mathbf{W}_n)$ is given by

$$(\mathbf{X}_n^t \mathbf{W}_n^{-1} \mathbf{X}_n)^{-1} \mathbf{X}_n^t \mathbf{W}_n^{-1} \boldsymbol{\Sigma}_n \mathbf{W}_n^{-1} \mathbf{X}_n (\mathbf{X}_n^t \mathbf{W}_n^{-1} \mathbf{X}_n)^{-1}$$

Thus, linear contrasts associated with $\hat{\boldsymbol{\beta}}(\mathbf{W}_n)$ will be consistent if both

$$\frac{1}{n}(\mathbf{X}_n^t \mathbf{W}_n^{-1} \mathbf{X}_n)$$

and

$$\frac{1}{n}(\mathbf{X}_n^t \mathbf{W}_n^{-1} \boldsymbol{\Sigma}_n \mathbf{W}_n^{-1} \mathbf{X}_n)$$

converge. These are both weighted covariance matrices, weighting subjects via the working covariance matrix in the first case and $\mathbf{W}_n^{-1}\boldsymbol{\Sigma}_n\mathbf{W}_n^{-1}$ in the latter. In the event that $\mathbf{W}_n = \mathbf{I}$ then the convergence of the former reduces to convergence of the variance covariance matrix of the regression variables. However, in more general cases, and the convergence of the latter weighted variance estimate, cannot be given without further restrictions. One setting where convergence can be obtained is where $\mathbf{W}_n$ and $\boldsymbol{\Sigma}_n$ have block diagonal structures as would be seen if one had repeated measurements on subjects.

In that case let n be the number of subjects and J be the number of observations within subjects. Further let: $\mathbf{X}_n^t = [\mathbf{Z}_1^t \ldots \mathbf{Z}_n^t]$, $\mathbf{W}_n = \mathbf{I}_n \otimes \mathbf{W}$ and $\boldsymbol{\Sigma}_n = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$ for $J \times p$ matrices $\mathbf{Z}_i$ and $J \times J$ matrices $\mathbf{W}$ and $\boldsymbol{\Sigma}$. Think of each $\mathbf{Z}_i$ as the covariates associated with the repeated measurements on subject i, $\boldsymbol{\Sigma}$ is the within subject correlation and $\mathbf{W}$ is our working version of the within subject correlation. Then our two necessary convergent series are:

$$\frac{1}{n}(\mathbf{X}_n^t \mathbf{W}_n^{-1} \mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^t \mathbf{W}^{-1} \mathbf{Z}_i,$$

and

$$\frac{1}{n}(\mathbf{X}_n^t \mathbf{W}_n^{-1} \boldsymbol{\Sigma}_n \mathbf{W}_n^{-1} \mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^t \mathbf{W}^{-1} \boldsymbol{\Sigma} \mathbf{W}^{-1} \mathbf{z}_i =$$

15.2 Normality

Ascertaining convergence to normality is a bit more involved. Fortunately, there's some convenient asymptotic theory to make life easier for us. Particularly, a version of the Central Limit Theorem states that if $E[\epsilon_i] = 0$ and Var$(\epsilon_i) = \sigma 2$ for ϵ_i iid and constants, $\mathbf{d}^t = [d_{n1} \dots d_{nn}]$ then

$$\frac{\sum_{i=1}^n d_{ni}\epsilon_i}{\sigma\sqrt{\sum_{i=1}^n d_{ni}^2}} = \frac{\mathbf{d}_n^t\boldsymbol{\epsilon}}{\sigma||\mathbf{d}_n||} \rightarrow N(0,1)$$

if $\max d_{ni}^2 = o(\sum_{i=1}^n d_{ni}^2)$.

With this theorem, we have all that we need. Let $\mathbf{Y} = \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Then note that

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = (\mathbf{X}_n^t\mathbf{X}_n)^{-1}\mathbf{X}_n^t\boldsymbol{\epsilon}.$$

And so,

$$\frac{\mathbf{q}^t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma\sqrt{\mathbf{q}^t(\mathbf{X}_n^t\mathbf{X}_n)^{-1}\mathbf{q}}} = \frac{\mathbf{q}^t(\mathbf{X}_n^t\mathbf{X}_n)^{-1}\mathbf{X}_n^t\boldsymbol{\epsilon}}{\sigma\sqrt{\mathbf{q}^t(\mathbf{X}_n^t\mathbf{X}_n)^{-1}\mathbf{q}}} = \frac{\mathbf{d}^t\boldsymbol{\epsilon}}{\sigma||\mathbf{d}||}$$

for $\mathbf{d} = \mathbf{q}^t(\mathbf{X}_n^t\mathbf{X}_n)^{-1}\mathbf{X}_n$. Thus, our linear contrast is $N(0,1)$, provided $\max \mathbf{q}^t(\mathbf{X}_n^t\mathbf{X}_n)^{-1}\mathbf{X}_n^t = o(\mathbf{q}^t(\mathbf{X}_n^t\mathbf{X}_n)^{-1}\mathbf{q})$. This will always be true if our $\mathbf{X}_n$ matrix is bounded.

Consider now our case where $\hat{\boldsymbol{\beta}}(\mathbf{w}_n) = (\mathbf{X}^t\mathbf{W}_n\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$. We assume that $\mathbf{Y}^t = [\mathbf{Y}_1^t \dots \mathbf{Y}_n^t]$, $\mathbf{X}^t = [\mathbf{Z}_1^t \dots \mathbf{Z}_n^t]$, $\mathbf{W}_n$ is a block matrix of $\mathbf{W}$ as assumed before. For context consider repeated measurements per subject. Let $\mathbf{Y}_i = \mathbf{Z}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ where Var$(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$. Then relying on our earlier work:

$$\begin{aligned}
\frac{\mathbf{q}^t(\hat{\boldsymbol{\beta}}(\mathbf{W}_n) - \boldsymbol{\beta})}{\text{SD}\{\mathbf{q}^t\hat{\boldsymbol{\beta}}(\mathbf{W}_n)\}} &= \frac{\mathbf{q}^t(\sum_{i=1}^n \mathbf{Z}_i^t\mathbf{W}\mathbf{z}_i)^{-1}\sum_{i=1}^n \mathbf{Z}_i^t\mathbf{W}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}_i}{\text{SD}\{\mathbf{q}^t\hat{\boldsymbol{\beta}}(\mathbf{W}_n)\}} \\
&= \frac{\mathbf{q}^t(\sum_{i=1}^n \mathbf{Z}_i^t\mathbf{W}\mathbf{z}_i)^{-1}\sum_{i=1}^n \mathbf{Z}_i^t\mathbf{W}\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{\epsilon}}_i}{\text{SD}\{\mathbf{q}^t\hat{\boldsymbol{\beta}}(\mathbf{W}_n)\}} \\
&= \frac{\sum_{i=1}^n \mathbf{d}_i\tilde{\boldsymbol{\epsilon}}_i}{\sqrt{\sum_{i=1}^n ||\mathbf{d}_i||^2}} \\
&= \frac{\sum_{i=1}^n ||\mathbf{d}_i||\frac{\mathbf{d}_i^t\tilde{\boldsymbol{\epsilon}}_i}{||\mathbf{d}_i||}}{\sqrt{\sum_{i=1}^n ||\mathbf{d}_i||^2}} \\
&= \frac{\sum_{i=1}^n ||\mathbf{d}_i||z_i}{\sqrt{\sum_{i=1}^n ||\mathbf{d}_i||^2}}
\end{aligned}$$

here $\tilde{\boldsymbol{\epsilon}}_i$ is $N(0, \mathbf{I}_J)$, z_i are iid with mean 0 and variance 1 and $\mathbf{d}_i = \mathbf{q}^t(\sum_{i=1}^n \mathbf{Z}_i^t\mathbf{W}\mathbf{z}_i)^{-1}\mathbf{Z}_i^t\mathbf{W}\boldsymbol{\Sigma}^{1/2}$. Thus we have reduced our statements down to the form of our generalized central limit theorem. Thus, we have shown that our estimates are asymptotically normal.

A final concern is that our statistic required $\boldsymbol{\Sigma}$. However, a consequence of Slutsky's theorem allows us to replace it with any consistent estimate. Let:

$$\mathbf{e}_i = \mathbf{Y}_i - (\mathbf{Z}_i^t\mathbf{Z}_i)^{-1}\mathbf{Z}_i^t\mathbf{Y}_i$$

then

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_i\mathbf{e}_i^t$$

is a consistent estimate of Σ. A last concern is the issue of assuming equal numbers of observations per subject. Generalizing this results in the same theory, just with more notational difficulties. (So we'll just assume that it's taken care of). Thus we have a fully formed methodology for performing inference on repeated measures data, where at no point did we presume knowledge of Σ, or even a good estimate of it. This form of analysis was later generalized into Generalized Estimating Equations (GEEs).

Chapter 16

Mixed models

It is often the case that parameters of interest in linear models are naturally thought of as being random rather than fixed. The rational for this can come about for many reasons. The first occurs when the natural asymptotics have the number of parameters tending to infinity with the sample size. As an example, consider the `Rail` dataset in `nlme`. The measurements are echo times for sound traveling along railroad lines (a measure of health of the line). Multiple (3) measurements are collected for each rail. Consider a model of the form

$$Y_{ij} = \mu + u_i + \epsilon_{ij},$$

where i is rail and j is measurement within rail. Treating the u_i as fixed effects results in a circumstance where the number of parameters goes to infinity with the rails. This can lead to inconsistent parameter estimates (Neyman and Scott, 1948) (for a simple example, see).

A solution to this problem is to put a distribution on the u_i, say $u_i \sim_{iid} N(0, \sigma_u^2)$. This is highly related to ridge regression (from the penalization chapter). However, unlike penalization, this problem allows for thinking about the random effect distribution as a population distribution (the population of rails in our example).

Perhaps the easiest way to think about random effects is to consider a fixed effect treatment of the u_i terms. Since we included an intercept, we would need to add one linear constraint on the u_i for identifiability. Consider the constraint, $\sum_{i=1}^n u_i = 0$. Then, μ would be interpreted as and overal mean and the u_i terms would be interpreted as the rail-specific deviation around that mean. The random effect model simply specifies that the U_i are iid $N(0, \sigma_u^2)$ and mutually independent from ϵ_{ij}. The mean of the distribution on the U_i has to be 0 (or fixed at a number), since it would not be identified from μ otherwise.

A perhaps preferable way to specify the model is hierarchically, $Y_{ij} \mid U_i \sim N(\mu, \sigma^2)$ and $U_i \mid \sim N(0, \sigma_U^2)$. Consider the impications of this model. First, note that

$$
\begin{aligned}
\text{Cov}(Y_{ij}, Y_{i'j'}) &= \text{Cov}(U_i + \epsilon_{ij}, U_{i'} + \epsilon_{i'j'}) &(16.1)\\
&= \text{Cov}(U_i, U_{i'}) + \text{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) &(16.2)\\
&= \left\{ \begin{array}{cc} \sigma^2 + \sigma_U^2 & \text{if } i = i'1 \text{ and } j = j' \\ \sigma^2 & \text{if } i = i' \text{ and } j \neq j' \\ 0 & \text{Otherwise} \end{array} \right. &(16.3)
\end{aligned}
$$

And thus the correlation between observations in the same cluster is $\sigma_u^2/(\sigma_u^2+\sigma^2)$. This is the ratio between the between subject variability, σ_u^2, and the total variability, $\sigma_u^2 + \sigma^2$.

Notice that the marginal model for $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^t$ is normally distributed with mean $\mu \times \mathbf{J}_{n_i}$ and variance $\sigma^2 \mathbf{I}_{n_i} + \mathbf{J}_{n_i}\mathbf{J}_{n_i}^t \sigma_u^2$. It is by maximizing this (marginal) likelihood that we obtain the ML estimates for $\mu, \sigma^2, \sigma_U^2$.

We can predict the U_i by considering the estimate $E[U_i \mid \mathbf{Y}]$. To derive this, note that the density for $U_i \mid \mathbf{Y}$ is equal to the density of $U_i \mid \mathbf{Y}_i$, since U_i is independent of every $Y_{i'j}$ for $i \neq i'$. Then further note that the density for $U_i \mid \mathbf{Y}_i$ is propotional to the joint density of Y_i, U_i, which is equal to the density of $Y_i \mid U_i$ times the density for U_i. Omitting anything that is not proportional in U_i, and taking twice the natural logarithm of the the densities, we obtain:

$$||\mathbf{Y}_i - \mu \mathbf{J}_{n_i} - U_i||^2/\sigma^2 + U_i/\sigma_U^2.$$

Expanding the square, and discarding terms that are constant in U_i, we obtain that U_i is normally distributed with mean

$$\frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma^2}{n}}(\bar{Y}_i - \mu).$$

Thus, if $\hat{\mu} = \bar{Y}$, our estimate of U_i is the estimate that we would typically use shrunken toward zero. The idea of shrinking estimates when simultaneously estimating several quantities is generally a good one. This has similarities with James/Stein estimation (see this review Efron and Morris, 1977).

Shrinkage estimation works by trading bias for lower variance. In our example, the shrinkage factor is $\sigma_u^2/(\sigma_u^2 + \sigma^2/n)$. Thus, the better estimated the mean for that group is (σ^2/n is small), or the more variable the group is (σ_u^2 is large), the less shrinkage we have. On the other hand, the fewer observations that we have, the larger the residual variation or the smaller the inter-subject variation, the more shrinkage we have. In this way the estimation is optimally calibrated to weigh the contribution of the individual versus the contribution of the group to the estimate regarding this specific individual.

16.1 General case

In general, we might write $\mathbf{Y} \mid \mathbf{U} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U}, \sigma^2\mathbf{I})$ and $\mathbf{U} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_U)$. This is marginally equivalent to specifying

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\epsilon}.$$

Here, the marginal likelihood for $\mathbf{Y}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^t + \sigma^2\mathbf{I}$. Maximum likelihood estimates maximize the marginal likelihood via direct numerical maximization or the EM algorithm (Dempster et al., 1977). Notice, for fixed variance components, the estimate of $\boldsymbol{\beta}$ is a weighted least squares estimate.

It should be noted the distinction between a mixed effect model and simply specifying a marginal variance structure. The same marginal likelihood could be obtained via the model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^t + \sigma^2\mathbf{I})$. However, some differences tend to arise. Often, the natural specification of a marginal variance structure doesn't impose positivity constraints that random effects do. For example, in the previous section, we saw that the covariance between measurements in the same cluster was $\sigma_u^2/(\sigma_u^2 + \sigma^2)$, which is guaranteed to be positive. However, if fitting a general marginal covariance structure, one would typically simply parameterize the covariance structure as either positive or negative.

Another difference lies in the hierarchical model itself. We can actually estimate the random effects if we specify them, unlike marginal models. This is a key (perhaps "the" key) defining attribute of mixed models. Again, our Best Linear Unbiased Predictor (BLUPs) is given by

$$E[bU \mid \mathbf{Y}]$$

As a homework exercise, derive the general form of the BLUPs

16.2 REML

Let $\mathbf{H}_X$ be the hat matrix for $\mathbf{X}$. Then note that

$$\mathbf{e} = (\mathbf{I} - \mathbf{H}_X)\mathbf{Y} = (\mathbf{I} - \mathbf{H}_X)\mathbf{Z}\mathbf{U} + (\mathbf{I} - \mathbf{H}_X)\boldsymbol{\epsilon}$$

Then, we can calculate the marginal distribution for $\boldsymbol{\epsilon}$ as singular normal with mean $(\mathbf{I} - \mathbf{H}_X)\mathbf{Z}\boldsymbol{\Sigma}_U\mathbf{Z}^t(\mathbf{I} - \mathbf{H}_X) + \sigma^2(\mathbf{I} - \mathbf{H}_X)$. Taking any full rank sub-vector of the $\boldsymbol{\epsilon}$ and maximizing the marginal likelihood for $\boldsymbol{\Sigma}_U$ and σ^2 is called restricted maximum likelihood (REML). REML estimates tend to be less biased than the ML estimates. For example, if $y_i \sim_{iid} N(\mu, \sigma^2)$, maximizing the likelihood for any $n - 1$ of the $e_i = y_i - \bar{y}$ yields the unbiased variance estimate (divided by $n - 1$) rather than the biased variance estimate obtained via maximum likelihood. REML estimates are often the default for linear mixed effect model programs.

An alternative way to derive the REML estimates is via Bayesian thinking. Consider a model where $\mathbf{Y} \mid \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^t + \sigma^2\mathbf{I})$ and $\boldsymbol{\beta} \sim N(0, \theta\mathbf{I})$. Calculating the mode for $\boldsymbol{\Sigma}_u$ and σ^2 after having integrated out $\boldsymbol{\beta}$ as $\theta \to \infty$ results in the REML estimates. While this is not terribly useful for general linear mixed effect modeling, it helps us think about REML as it relates to Bayesian analysis and it allows us to extend REML in settings where residuals are less well defined, like generalized linear mixed models.

16.3 Prediction

Consider generally trying to predict U from observed data Y. Let f_{uy}, f_u, f_y, $f_{u|y}$ and $f_{y|u}$ be the joint, marginal and conditional densities respectively. Let $\theta(Y)$ be our estimator of U. Consider evaluating the prediction error via the expected squared loss

$$E[(U - \theta(Y))^2]$$

We now show that this is minimized at $\theta(Y) = E[U \mid Y]$. Note that

$$\begin{aligned}
&E[(U - \theta(Y))^2] \\
&= E[(U - E[U \mid Y] + E[U \mid Y] - \theta(Y))^2] \\
&= E[(U - E[U \mid Y])^2] - 2E[(U - E[U \mid Y])]E[(E[U \mid Y] - \theta(Y))] + E[(E[U \mid Y] - \theta(Y))^2] \\
&= E[(U - E[U \mid Y])^2] + E[(E[U \mid Y] - \theta(Y))^2] \\
&\geq E[(U - E[U \mid Y])^2].
\end{aligned}$$

This argument should seem familiar. (In fact, Hilbert space results generalize these kinds of arguments into one theorem.) Therefore, $E[U \mid Y]$ is the best predictor. Note, that it is always the best predictor, regardless of the settting. Furthermore, in the context of linear models, this predictor is both linear (in $\mathbf{Y}$) and unbiased. We mean unbiased in the sense of:

$$E[U - E[U \mid Y]] = 0.$$

Therefore, even in the more restricted class of linear estimators, in the case of mixed models, $E[U \mid Y]$ remains best.

A complication arises in that we do not know the variance components. As that is the case, we must plug in the estimates (either REML or ML). The BLUPs lose their optimality properties then and are thus often called EBLUPs (for empirical BLUPs).

Prediction of this sort relates to so-called empirical Bayesian prediction and shrinkage estimation. In your more advanced classes on decision theory, you'll learn about loss functions and uniform desirability of shrinkage estimators over the straightforward estimators. (In our case the straightforward estimator is the one that treats the random effects as if fixed.) This line of thinking yields yet another use for random effect models, where we might apply them merely for the benefits of shrinkage, but don't actually think of our random effects as if random. Consider settings like genomics. The genes being studied are exactly the quantities of interest, not a random sample from a population of genes. However, it remains useful to treat effects associated with genes as if random to obtain the benefits of shrinkage.

16.4 P-splines

16.4.1 Regression splines

The application to splines has been a very successful, relatively new, use of mixed models. To discuss the methodology, we need to introduce splines briefly. We will only overview this area and focus on regression splines, while acknowledging that other spline bases may be preferable.

Let $(a)_+ = a$ if $a > 0$ and 0 otherwise. Let ξ be a known knot location. Now consider the model:

$$E[Y_i] = \beta_0 + \beta_1 x_i + \gamma_1 (x_i - \xi)_+.$$

For x_i values less than or equal to ξ, we have

$$E[Y_i] = \beta_0 + \beta_1 x_i$$

and for x_i values above ξ we have

$$E[Y_i] = (\beta_0 + \gamma_1 x i) + (\beta_1 + \gamma_1) x_i.$$

Thus, the response function $f(x) = \beta_0 = \beta_1 x + \gamma_1 (x - \xi)_+$ is continuous at ξ and is a line before and after. This allows us to create ”hockey stick” models, with a line below ξ and a line with a different slope afterwards. Furthermore, we could expand the model as

$$E[Y_i] = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \gamma_k (x_i - \xi_k)_+.$$

where ξ_k are knot points. Not the model is a spiky, but flexible, function that is linear between the knots and meets at the knots.

To make the fit less spiky, we want continuous differentiability at the knot points. First note that the function $(x)_+^p$ has $p - 1$ continuous derivatives at 0. To see this, take the limit to zero of the derivatives from the right and the left. Thus, the function

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{k=1}^{K} \gamma_k (x_i - \xi_k)_+^2$$

will consist of parabolas between the knot points and will have one continuous derivative at the knot points. This will fit a smooth function that can accommodate a wide variety of data shapes.

16.4.2 Coding example

16.5 Further reading

A great book on mixed models in R is Pinheiro and Bates (2006). In addition, the books by Searle, McCulloch, Verbeke and Molenberghs are wonderful treatments of the topic (McCulloch and Searle, 2001; Verbeke and Molenberghs, 2009). Finally, the newer package `lme4` has a series of vignettes.

Chapter 17

Bayes analysis

17.1 Introduction to Bayesian analysis

Bayesian analysis is a form of statistical inference relying on Baye's rule. The general version of Baye's rule states that

$$f(y|x) = f(x|y)f(y)/f(x)$$

where we're using f (loosely) as the appropriate density, mass function or probability and x and y represent random variables or events.

In the context of Bayesian analysis, Baye's rule is used in the following way. Let $\mathcal{L}(\theta; y)$ be the likelihood associated with data, y, and parameter θ. We codify our prior knowledge about θ with a prior distribution, $\pi(\theta)$. Then, a Bayesian analysis is performed via the posterior distribution

$$\pi(\theta \mid y) = f(y \mid \theta)\pi(\theta)/f(y) \propto_\theta f(y \mid \theta)\pi(\theta) \propto_\theta \mathcal{L}(\theta; y)\pi(\theta).$$

Therefore, one obtains the posterior, up to multiplicative constants, by multiplying the likelihood times the prior.

Coupled with Bayesian analysis is Bayesian interpretation of the probabilities. The prior is viewed a belief and the posterior is then an updated belief coupling the objective evidence (the likelihood) with the subjective belief (the prior). By viewing probabilities as personal quantifications of beliefs, a Bayesian can talk about the probability of things that frequentists cant. So, for example, if I roll a die and don't show you the result, you as a Bayesian can say that the probability that this specific roll is a six is one sixth. You as a frequentist, in contrast, must say that in one sixth of repetitions of this experiment, the result will be a six. To a frequentist, this specific roll is either six or not.

This distinction in probabilistic interpretation has consequences in statistical interpretations. For example in diagnostic tests, a Bayesian can talk about the probability that a person has a disease, whereas a frequency interpretation relies on the percentage of diseased people in a population of those similar.

Personally, I've never minded either interpretation, but to many, the Bayesian interpretation seems more natural. In contrast, many practitioners dislike Bayesian analysis because of the prior specification, and the heavy reliance on fully specified models.

It should be noted that the discussion up to this point contrasted classical frequency thinking with classical subjective Bayesian thinking. In fact, most modern applied statisticians use hybrid approaches. They might, for example, develop a procedure with Bayesian tools (the manipulation of conditional distributions with priors on the parameters), but evaluate the procedure using frequency error rates. For all intents and purposes, such a procedure is frequentist, just developed with a Bayesian mindset. In contrast, many frequency statistical practitioners interpret their results with an approximate Bayesian mindset. Such procedures are simply Bayesian without the formalism. Even between these approaches there's continuous shades of gray. Therefore, saying a modern statistician is either Bayesian or frequentist is usually misleading, unless that person does research or writes on statistical foundations. Nonetheless, foundational thinking is useful for understanding and clarifying thinking. It's worth then reading and internatlizing the literature on foundations for this reason alone. It's a lot like working on core drills to get better at a sport. That and it's quite a bit of fun! Some of my favorite modern writers on the topic include (heavily emphasizing people I know pretty well or have run into recently): Jim Berger, Nancy Reid, Richard Royall, Deborah Mayo, David Cox, Charles Rohde, Andrew Gelman, Larry Wasserman and Jeff Blume. Their work will point to many others (Basu, Birnbaum, De Finetti, Lindley all also come to mind).

Finally, it should also be noted that Bayes versus frequency is far from the only schism in statistical foundations. Personally, I find the distinction between direct use of the design in frequency analysis to obtain robustness, like is often done is randomization testing and survey sampling, versus fully specified modeling a larger distinction than how one uses the model (Bayes versus frequentist). In addition, causal analysis versus association (non-causal) analysis forms a large distinction and one can perform Bayes or frequentists causal analysis and non-causal analyses. Furthermore, the likelihood paradigm (Royall, 1997) offers a third inferential technique given a model over Bayes and frequency interpretations.

17.2 Basic Bayesian models

17.2.1 Binomial

We'll begin our discussion of Bayesian models by using some count outcome cases to build intuition. First, consider a series of coin flips, $X_1, \ldots, X_n \sim$ Bernoulli(θ). The likelihood associated with this experiment is

$$\mathcal{L}(\theta) \propto \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} = \theta^x(1-\theta)^{n-x}$$

where $x = \sum_i x_i$. Notice the likelihood depends only on the total number of successes. Consider putting a Beta(α, β) prior on θ. The the posterior is

$$\pi(\theta \mid x) \propto_\theta \mathcal{L}(\theta) \times \pi(\theta) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

therefore the posterior distribution is Beta$(x+\alpha, n-x+\beta)$. The posterior mean is

$$E[\theta \mid x] = \frac{x+\alpha}{n+\alpha+\beta} = \delta\hat{p} + (1-\delta)\frac{\alpha}{\alpha+\beta}$$

Therefore, the posterior mean is a weighted average of the MLE ($\hat{p}$) and the prior mean $\frac{\alpha}{\alpha+\beta}$. The weight is

$$\delta = \frac{n}{n+\alpha+\beta}.$$

Notice that, as $n \to \infty$ for fixed α and β, $\delta \to 1$ and the MLE dominates. That is, as we collect more data, the prior becomes less relevant and the data, in the form of the likelihood, dominates. On the other hand, for fixed n, as either α or β go to infinity (or both), the prior dominates ($\delta \to 0$). For the Beta distribution α or β going to infinity makes the distribution much more peaked. Thus, if we are more certain of our prior distribution, the data matters less.

17.2.2 Poisson

Let $X \sim$ Poisson$(t\lambda)$. Then

$$\mathcal{L}(\lambda) \propto \lambda^x e^{-t\lambda}.$$

Consider putting a Gamma(α, τ^{-1}) prior on λ. Then we have that

$$\pi(\lambda \mid x) \propto \lambda^{x+\alpha-1} e^{-\lambda(t+\tau)}$$

and thus the posterior is Gamma$(x+\alpha, (t+\tau)^{-1})$. Because of the inversion of the second scale parameter of the Gamma, often Bayesians specify it in the terms of the inverse (as in Gamma$(x+\alpha, t+\tau)$). Often to avoid confusion, the mean of the gamma will be given to ensure no confusion over the parameterization.

The posterior mean is:

$$E[\lambda \mid x] = \frac{x+\alpha}{t+\tau} = \delta\hat{\lambda} + (1-\delta)\frac{\alpha}{\tau}$$

where $\hat{\lambda} = x/t$ is the MLE (the observed rate) and α/τ is the prior estimate. In this case

$$\delta = \frac{t}{t+\tau}$$

so that as $t \to \infty$ the MLE dominates while the prior dominates as $\tau \to \infty$.

17.3 Bayesian Linear Models

17.3.1 Bayesian Linear Models

Recall our standard Gaussian linear model, where $\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Consider three common prior specifications:

1. $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Sigma}_0)$ and $\sigma^{-2} \sim$ Gamma(α_0, τ_0^{-1}).
2. $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ and $\sigma^{-2} \sim$ Gamma(α_0, τ_0^{-1}).

3. $(\boldsymbol{\beta}, \sigma^2) \sim \sigma^{-2}$.

The final prior specification is not a proper density. It doesn't have a defined integral for the elements of $\boldsymbol{\beta}$ from $-\infty$ to ∞ and for $0 \leq \sigma^2 < \infty$. However, proceeding as if it were a proper density yields a proper distribution for the posterior. Such "improper" priors are often used to specify putatively uninformative distributions that yield valid posteriors. In this case, the posterior has the property of the posterior mode being centered around $\hat{\boldsymbol{\beta}}$.

The distinction between the first case and the second is the inclusion of σ^2 in the prior specification for $\boldsymbol{\beta}$. This is useful for making all posterior distributions tractable, including that of $\boldsymbol{\beta}$ integrated over σ^2. However, it may or may not reflect the desired prior distribution.

The second specification has tractable full conditionals. That is, we can easily figure out $\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}$ and $\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}$. However, the posterior marginals of the parameters ($\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}$ in particular) are not tractable. This posterior is often explored using Monte Carlo.

The third specification is also completely tractable.

17.4 Monte Carlo sampling

Even though many of our Bayesian models are completely tractable, we will explore the posteriors via Monte Carlo. The reason for this is to get students familiar with Monte Carlo so that they can apply it in the more complex settings that they are likely to encounter in practice. Specifically, usually fully tractable posteriors are more of an exception than the rule. For the most part, for linear models, one should use the fully tractable results as they are much faster.

We now give some strategies for Monte Carlo sampling from a posterior.

17.4.1 Sequential Monte Carlo

Notice for three variables, X, Y and Z, sampling $f_z(z)$, $f_{y|z}(y)$ and $f_{x|y,z}(x \mid y, z)$ yields a multivariate draw from the joint distribution $f(x, y, z)$. So, for example, consider setting 3 of our prior specifications

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X} \sim N(\hat{\boldsymbol{\beta}}, \mathbf{X}^t\mathbf{X}\sigma^2) \text{ and } \sigma^{-2} \mid \mathbf{y}, \mathbf{X} \sim \text{Gamma}\{(n-p)/2, 2/(n-p)S^2\}$$

Notice that $E[\sigma^{-2} \mid \mathbf{X}, \mathbf{y}] = 1/S^2$. To simualte from this distribution, we first simulate from $\sigma^{-2} \mid \mathbf{X}, \mathbf{y}$ then plug that simulation into $\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}$ and simulate an $\boldsymbol{\beta}$. The pair is a draw from the joint posterior distribution of $\boldsymbol{\beta}$ and σ^{-2}.

17.4.2 Gibbs sampling

Consider again our three random variables. Suppose that an initial value of x and y, say $x^(1)$ and $y^{(1)}$ was obtained. Then consider simulating

1. $z^{(1)} \sim f_{z|x,y}(z \mid x^{(1)}, y^{(1)})$
2. $x^{(2)} \sim f_{x|y,z}(x \mid y^{(1)}, z^{(2)})$
3. $y^{(2)} \sim f_{y|x,z}(y \mid x^{(2)}, z^{(2)})$
4. $z^{(2)} \sim f_{z|x,y}(z \mid x^{(2)}, y^{(2)})$
5. $x^{(3)} \sim f_{x|y,z}(x \mid y^{(2)}, z^{(3)})$

and so on. In other words, always update a simulated variable using the most recently simulated version of the other variables. In fact, one need not use the full conditionals. Any collection of conditionals would work. Moreover, any random order works, or even randomizing the order each iteration. However, some conditions have to be met for the asymptotics of the sampler to work. For example, you have to update every variable infinitely often and the whole space has to be explorable by the sampler. If the conditions are met, the sampler is a Markov chain whose stationary distribution, i.e. the limiting distribution, is $f(x, y, z)$. Moreover, there's lots of results saying that you can use the ouput of the sampler in much the same way one uses iid samples. For example approximating posterior means with the average of the simulated variables. However, the Markovian nature of the sampler makes using the samples a little trickier. One could try to combat this by running the chain for a while and throwing out all of the early simulations used to "burn in" the sampler. This throws away a lot of data. Our preferred method is to use good starting values (why not start at the MLE?) and use all of the simulated data.

Let's illustrate the sampler with prior specification prior 2. Consider the simplified model:

$$\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{X}, \theta \sim N(\mathbf{X}\boldsymbol{\beta}, \theta^{-1}\mathbf{I}) \text{ and } \boldsymbol{\beta} \sim N(\mathbf{0}, \psi^{-1}\mathbf{I}) \text{ and } \theta \sim \text{Gamma}(\alpha/2, \tau^{-1}.$$

Under this specification, the full conditionals are:

$$\begin{aligned} \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \theta &\sim N\{(\mathbf{X}^t\mathbf{X}\theta + \psi\mathbf{I})^{-1}\mathbf{X}^t\mathbf{y}, (\mathbf{X}^t\mathbf{X}\theta + \psi\mathbf{I})^{-1}\} \\ \theta \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\beta} &\sim \text{Gamma}\{(n+\alpha)/2, 2(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau)^{-1}\} \end{aligned}$$

```
data("mtcars")
y = mtcars$mpg - mean(mtcars$mpg)
x = cbind(1, mtcars$wt - mean(mtcars$wt))

n = length(y)
p = ncol(x)

fitML  = lm(y ~ x - 1)


xtx = t(x) %*% x
xty = as.vector(t(x) %*% y)
```

```
nosim = 10000

rmvnorm = function(mu, Sigma) as.vector(mu + chol(Sigma) %*% rnorm(length(mu)))

thetaCurrent = 1 / summary(fitML)$sigma^2
beta = NULL
theta = thetaCurrent * 100
psi = .01
alpha = .01
tau = .01 * summary(fitML)$sigma^2
for (i in 1 : nosim){
  V = solve(xtx * thetaCurrent + psi * diag(1, p, p))
  mu = V %*% xty * thetaCurrent
  betaCurrent = rmvnorm(mu, V)
  sumesq = sum((y - x %*% betaCurrent)^2)
  thetaCurrent = rgamma(1, (n + alpha) / 2, rate = (sumesq + tau)/2)
  theta = c(theta, thetaCurrent)
  beta = rbind(beta, betaCurrent)
}
sigma = sqrt(1/ theta)
quantile(beta[,1], c(.025, .975))
quantile(beta2[,2], c(.025, .975))
quantile(sigma, c(0.025, .975))
```

Bibliography

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Efron, B. and Morris, C. N. (1977). *Stein's paradox in statistics*. WH Freeman.

McCulloch, C. E. and Searle, S. R. (2001). Linear mixed models (lmms). *Generalized, linear, and mixed models*, pages 156–186.

Myers, R. H. (1990). *Classical and modern regression with applications*, volume 2. Duxbury Press Belmont, CA.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*, volume 71. CRC press.

Searle, S. R. (2012). *Linear models*. Wiley.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.