Introduction to statistics: Matrix formulatuon of the linear model

Shravan Vasishth

Universität Potsdam
vasishth@uni-potsdam.de
http://www.ling.uni-potsdam.de/∼vasishth

April 12, 2020

Matrix form of the linear model

- This lecture is not crucial for following this course.
- It is however useful to have seen the matrix form of the linear model once.
- Having see this formulation once will help in understanding contrast coding.

The linear model

Consider a deterministic function $\phi(\mathbf{x}, \beta)$ which takes as input some variable values x and some fixed values β. A simple (if abstract) example would be

$$y = \beta x \tag{1}$$

The linear model

Another example with two fixed values β_0 and β_1 is:

$$y = \beta_0 + \beta_1 x \tag{2}$$

We can rewrite the above equation as follows.

The linear model

$$
\begin{aligned}
y =& \beta_0 + \beta_1 x \\
=& \beta_0 \times 1 + \beta_1 x \\
=& \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\
y =& \phi(x, \beta)
\end{aligned} \tag{3}
$$

The linear model

In a statistical model, we don't expect an equation like $y = \phi(x, \beta)$ to fit all the points exactly.
For example, we could come up with an equation that, given a person's weight, gives a prediction of their height:

$$\text{predicted height} = \beta_0 + \beta_1 \text{weight} \tag{4}$$

The linear model

Given any single value of the weight of a person, we will probably not get a perfectly correct prediction of the height of that person. This leads us to a non-deterministic version of the above function:

$$y = \phi(x, \beta, \varepsilon) = \beta_0 + \beta_1 x + \varepsilon \tag{5}$$

The linear model

Here, ε is an error random variable which is assumed to have some PDF (the normal distribution) associated with it. It is assumed to have expectation (mean) 0, and some standard deviation (to be estimated from the data) σ.
We can write this statement in compact form as $\varepsilon \sim N(0, \sigma)$.
The **general linear model** is a non-deterministic function like the one above:

$$Y = f(x)^T \beta + \varepsilon \tag{6}$$

The linear model

The matrix formulation will be written as below. n refers to the number of data points (that is, $Y_1, \ldots, Y_n$), and the index j ranges from 1 to n.

$$Y = X\beta + \varepsilon \Leftrightarrow y_j = f(x_j)^T\beta + \varepsilon_j, j = 1, \ldots, n \tag{7}$$

The linear model

To make this concrete, suppose we have three data points, i.e., n=3. Then, the matrix formulation is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \varepsilon \tag{8}$$

$$Y = X\beta + \varepsilon \tag{9}$$

Here, $f(x_1)^T = (1\ x_1)$, and is the first row of the matrix X, $f(x_2)^T = (1\ x_2)$ is the second row, and $f(x_3)^T = (1\ x_3)$ is the third row.

The linear model

Note that $E[Y] = X\beta$. β is a $p \times 1$ matrix, and X, the **design matrix**, is $n \times p$.

Geometric argument

- When we have a deterministic model $y = \phi(f(x)^T, \beta) = \beta_0 + \beta_1 x$, this implies a perfect fit to all data points.
- This is like solving the equation $Ax = b$ in linear algebra: we solve for β in $X\beta = y$ using, e.g., Gaussian elimination.
- When we have a non-deterministic model $y = \phi(f(x)^T, \beta, \varepsilon)$, there is no solution. Now, the best we can do is to get Ax to be as close an approximation as possible to b in $Ax = b$.
- In other words, we try to minimize $\mid b - Ax \mid$.

Geometric argument

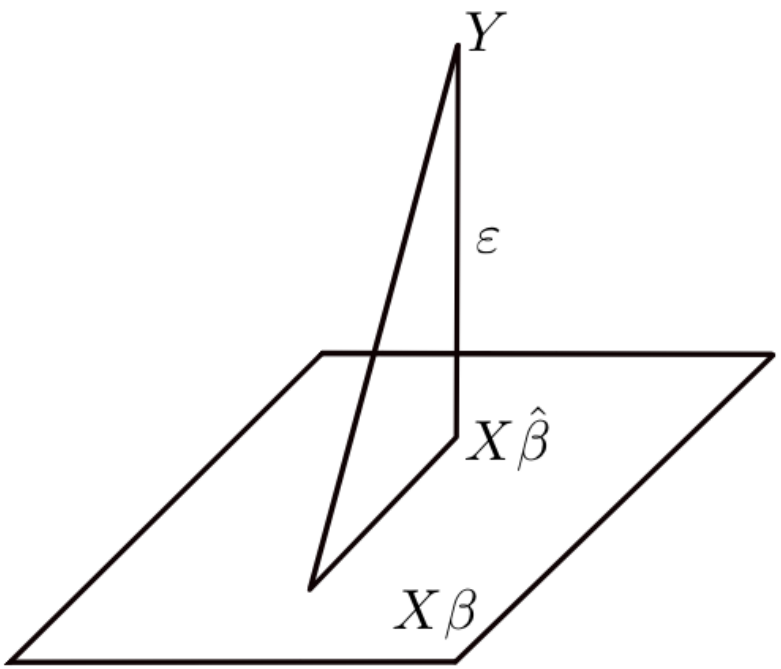
- The goal is to estimate β; we want to find a value of β such that the observed Y is as close to its expected value $X\beta$.
- In order to be able to identify β from $X\beta$, the linear transformation $\beta \rightarrow X\beta$ should be one-to-one, so that every possible value of β gives a different $X\beta$.

This in turn requires that X be of full rank p.
If X is an $n \times p$ matrix, then it is necessary that $n \geq p$. There must be at least as many observations as parameters. If this is not true, then the model is said to be **over-parameterized**.

Geometric argument

- Assuming that X is of full rank, and that $n > p$, Y can be considered a point in n-dimensional space and the set of candidate $X\beta$ is a p-dimensional subspace of this space; see the figure on the next slide.
- There will be one point in this subspace which is closest to Y in terms of Euclidean distance. The unique β that corresponds to this point is the **least squares estimator** of β; we will call this estimator $\hat{\beta}$.

Geometric argument



Geometric argument

Notice that $\varepsilon = (Y - X\hat{\beta})$ and $X\beta$ are perpendicular to each other. Because the dot product of two perpendicular (orthogonal) vectors is 0, we get the result:

$$(Y - X\hat{\beta})^T X\beta = 0 \Leftrightarrow (Y - X\hat{\beta})^T X = 0 \tag{10}$$

Example:

```
x1<-c(1,0)
x2<-c(0,1)
sum(x1*x2)

## [1] 0
```

Geometric argument

Multiplying out the terms, we proceed as follows. One result that we use here is that $(AB)^T = B^T A^T$.

$$
\begin{aligned}
&(Y - X\hat{\beta})^T X = 0 \\
&(Y^T - \hat{\beta}^T X^T) X = 0 \\
\Leftrightarrow &Y^T X - \hat{\beta}^T X^T X = 0 \\
\Leftrightarrow &Y^T X = \hat{\beta}^T X^T X \\
\Leftrightarrow &(Y^T X)^T = (\hat{\beta}^T X^T X)^T \\
\Leftrightarrow &X^T Y = X^T X \hat{\beta}
\end{aligned}
\tag{11}
$$

Geometric argument

This gives us the important result:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{12}$$

X is of full rank, therefore $X^T X$ is invertible.

- $X^+ = (X^T X)^{-1} X^T$ is the **generalized matrix inverse** of the design matrix X, and will become very important for us in the **contrast coding** lecture.
- We will see in the contrast coding lecture how X^+ defines hypothesis tests in the linear (mixed) model.

Geometric argument

Example:

```
(X<-matrix(c(rep(1,8),rep(c(-1,1),each=4),
             rep(c(-1,1),each=2,2)),ncol=3))

##      [,1] [,2] [,3]
## [1,]    1   -1   -1
## [2,]    1   -1   -1
## [3,]    1   -1    1
## [4,]    1   -1    1
## [5,]    1    1   -1
## [6,]    1    1   -1
## [7,]    1    1    1
## [8,]    1    1    1
```

Geometric argument

```
library(Matrix)
## full rank:
rankMatrix(X)

## [1] 3
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 1.7764e-15
```

The mean and variance of the parameters beta

Our model is:

$$Y = X\beta + \varepsilon \tag{13}$$

The mean and variance of the parameters beta

Let $\varepsilon \sim N(0, \sigma^2)$.
In other words, we are assuming that each value generated by the random variable ε is independent and it has the same distribution, i.e., it is identically distributed. This is sometimes shortened to the iid assumption.
So we should technically be writing:

$$Y = X\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma) \tag{14}$$

and add that Y are independent and identically distributed.
We could also have written:

$$Y \sim Normal(X\beta, \sigma) \tag{15}$$

The mean and variance of the parameters beta

Some consequences of the above statements:

1. $E(\varepsilon) = 0$
2. $Var(\varepsilon) = \sigma^2 I_n$
3. $E[Y] = X\beta = \mu$
4. $Var(Y) = \sigma^2 I_n$

I_n here is the identity matrix, e.g.:

```
I<-matrix(c(1,0,0,1),ncol = 2)
sigma<-10
sigma*I
```

```
##      [,1] [,2]
## [1,]   10    0
## [2,]    0   10
```

The mean and variance of the parameters beta

We can now derive the expectation and variance of the estimators $\hat{\beta}$.

We need a fact about variances: $Var(aB)$, where a is a constant, is $a^2 Var(B)$. In the matrix setting, Var(AB), where A is a constant, is $AVar(B)A^T$.

$$E[\hat{\beta}] = E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T X \beta = \beta \qquad (16)$$

Notice that the above shows that $\hat{\beta}$ is an unbiased estimator of β.

The mean and variance of the parameters beta

Next, we compute the variance:

$$Var(\hat{\beta}) = Var([(X^T X)^{-1} X^T]Y) \tag{17}$$

Expanding the right hand side out:

$$Var([(X^T X)^{-1} X^T]Y) = [(X^T X)^{-1} X^T]Var(Y)[(X^T X)^{-1} X^T]^T \tag{18}$$

The mean and variance of the parameters beta

Replacing Var(Y) with its variance $\sigma^2 I$, and unpacking the transpose on the right-most expression $[(X^T X)^{-1} X^T]^T$:

$$Var(\hat{\beta}) = [(X^T X)^{-1} X^T] \sigma^2 I X [(X^T X)^{-1}]^T \tag{19}$$

Since σ^2 is a scalar we can move it to the left, and any matrix multiplied by I is the matrix itself, so we ignore I, getting:

$$Var(\hat{\beta}) = \sigma^2 [(X^T X)^{-1} X^T X [(X^T X)^{-1}]^T \tag{20}$$

The mean and variance of the parameters beta

Since $(X^TX)^{-1}X^TX = I$, we can simplify to

$$Var(\hat{\beta}) = \sigma^2[(X^TX)^{-1}]^T \tag{21}$$

Now, $(X^TX)^{-1}$ is symmetric, so $[(X^TX)^{-1}]^T = (X^TX)^{-1}$. This gives us:

$$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1} \tag{22}$$

The mean and variance of the parameters beta

An example:

```
hindi<-read.table("data/hindiJEMR.txt",header=TRUE)
y<-as.matrix(hindi$TFT)
x<-scale(log(hindi$word_len),scale=FALSE)
m0<-lm(y~x)
```

The mean and variance of the parameters beta

```
summary(m0)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -485   -162    -31     85   5566
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  246.892      0.848     291   <2e-16
## x            189.429      1.602     118   <2e-16
##
## Residual standard error: 240 on 79941 degrees of freedom
## Multiple R-squared:  0.149,Adjusted R-squared:  0.149
## F-statistic: 1.4e+04 on 1 and 79941 DF,  p-value: <2e-16
```

The mean and variance of the parameters beta

```
## design matrix:
X<-model.matrix(m0)
head(X,n=4)

##   (Intercept)         x
## 1           1  0.46907
## 2           1 -0.44722
## 3           1 -0.44722
## 4           1  0.46907
```

The mean and variance of the parameters beta

```
## (X^TX)^{-1}
invXTX<-solve(t(X)%*%X)
## estimate of beta:
(beta<-invXTX%*%t(X)%*%y)

##                [,1]
## (Intercept) 246.89
## x           189.43
```

The mean and variance of the parameters beta

```
## estimated SD (se) of the estimate of beta:
(hat_sigma<-summary(m0)$sigma)

## [1] 239.88

(hat_var<-hat_sigma^2*invXTX)

##               (Intercept)             x
## (Intercept)  7.1979e-01 -1.3534e-14
## x           -1.3534e-14  2.5679e+00
```

The mean and variance of the parameters beta

What we have here is a bivariate normal distribution as an estimate of the β parameters:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N(\begin{pmatrix} 246.89165 \\ 189.42927 \end{pmatrix}, \begin{pmatrix} 0.71979 & -1.35342 \times 10^{-14} \\ -1.35342 \times 10^{-14} & 2.56791 \end{pmatrix}) \tag{23}$$

The mean and variance of the parameters beta

The variance of a bivariate distribution has the variances along the diagonal, and the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ on the off-diagonals. Covariance is defined as:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \rho \sigma_{\hat{\beta}_0} \sigma_{\hat{\beta}_1} \tag{24}$$

where ρ is the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$.

The mean and variance of the parameters beta

So $\hat{\beta}_0 \sim N(\mu = 246.89165, \sigma^2 = 0.71979)$ and $\hat{\beta}_1 \sim N(189.42927, \sigma^2 = 2.56791)$, and $Cov(\beta_0, \beta_1) = -1.35342 \times 10^{-14}$. So the correlation between the $\hat{\beta}$ is

```
## hat rho:
hat_var[1,2]/(sqrt(hat_var[1,1])*sqrt(hat_var[2,2]))

## [1] -9.955e-15
```

This is the correlation between the sampling distributions of β_0 and β_1.

Exercise: The effect of centering

Above, we fit the model after centering the predictor:

```
y<-as.matrix(hindi$TFT)
## centering:
x<-scale(log(hindi$word_len),scale=FALSE)
m0<-lm(y~x)
```

Redo this model with x uncentered and unscaled:

```
y<-as.matrix(hindi$TFT)
## no centering
x<-log(hindi$word_len)
m0<-lm(y~x)
```

What is the correlation between the sampling distributions of the intercept and slope now?

The likelihood ratio test

Define the likelihood ratio statistic as:

$$\Lambda = \frac{max_{\theta \in \omega_0}(lik(\theta))}{max_{\theta \in \omega_1)}(lik(\theta))} \tag{25}$$

where, $\omega_0 = \{\mu_0\}$ and $\omega_1 = \{\forall \mu \mid \mu \neq \mu_0\}$.
Suppose that $X_1, \ldots, X_n$ are iid and normally distributed with mean μ and standard deviation σ (assume for simplicity that σ is known).

The likelihood ratio test

Let the null hypothesis be $H_0 : \mu = \mu_0$ and the alternative be $H_1 : \mu \neq \mu_0$. Here, μ_0 is a number, such as 0. Now, the numerator of the likelihood ratio statistic is:

$$\frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right) \tag{26}$$

For the denominator, the maximum likelihood can be achieved by specifying the MLE $\bar{X}$ as μ:

$$\frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \tag{27}$$

The likelihood ratio test

The likelihood ratio statistic is then:

$$\Lambda = \frac{\frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)}{\frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)} \tag{28}$$

The likelihood ratio test

Canceling out common terms:

$$\Lambda = \frac{exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)}{exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)} \tag{29}$$

The likelihood ratio test

Taking logs:

$$\begin{aligned}\log \Lambda &= \left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right) - \left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) \\ &= -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X})^2\right)\end{aligned} \tag{30}$$

The likelihood ratio test

Now, note that

$$\sum_{i=1}^{n}(X_i - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 \qquad (31)$$

The likelihood ratio test

This means that

$$\sum_{i=1}^{n}(X_i - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \bar{X})^2 = n(\bar{X} - \mu_0)^2 \qquad (32)$$

The likelihood ratio test

So, we can write $\log \Lambda = \ell$ as:

$$\ell = -\frac{1}{2\sigma^2} n(\bar{X} - \mu_0)^2 \tag{33}$$

Rearranging terms:

$$-2\ell = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} \tag{34}$$

The likelihood ratio test

Or maybe even more transparently:

$$-2\ell = \frac{(\bar{X} - \mu_0)^2}{\frac{\sigma^2}{n}} \tag{35}$$

This should remind you of the t-test! Basically, all this is saying is that we reject the null when $\mid \bar{X} - \mu_0 \mid$ is large.

The likelihood ratio test

More generally, we will define the **likelihood ratio test statistic** as follows. Here, $\Lambda(\theta)$ refers to the likelihood given some value of θ, and $\ell(\theta)$ or $\log \Lambda$ refers to the log likelihood.
Δ is used to stress the fact that we are talking about the ratio (or difference, in log likelihood) of likelihood. We could have just used Λ or ℓ.

$$\begin{aligned} \Delta\Lambda &= -2 \times \log(\Lambda(\theta_0)/L(\theta_1)) \\ \Delta \log \Lambda &= -2 \times \{\ell(\theta_0) - \ell(\theta_1)\} \\ \Delta \log \Lambda &= -2 \times \{\ell(\theta_0) - \ell(\theta_1)\} \end{aligned} \tag{36}$$

where θ_1 and θ_0 are the estimates of θ under the alternative and null hypotheses, respectively. The likelihood ratio test rejects H_0 if $\Delta \log \Lambda$ is sufficiently large. As the sample size approaches infinity:

The likelihood ratio test

This is called Wilks' theorem. The proof of Wilks' theorem is fairly involved but you can find it on the internet if you are interested, or in Lehmann's *Testing Statistical Hypotheses*.

The likelihood ratio test

Note that sometimes you will see the form:

$$\Delta \log \Lambda = 2\{\ell(\theta_1) - \ell(\theta_0)\} \tag{38}$$

It should be clear that both statements are saying the same thing; in the second case, we are just subtracting the null hypothesis log likelihood from the alternative hypothesis log likelihood.

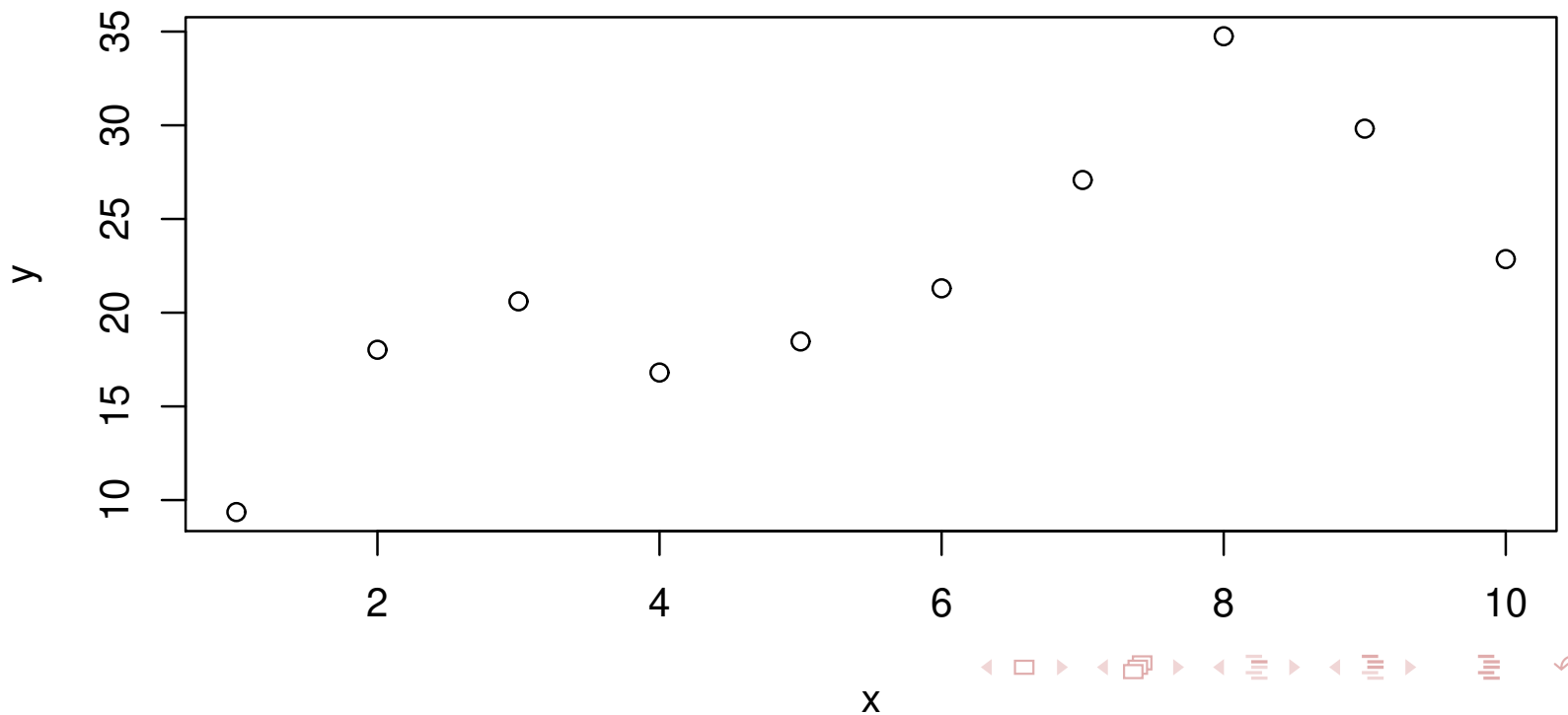
The likelihood ratio test

A practical example will make the usage of this test clear. Let's just simulate a linear model:

```
x<-1:10
y<- 10 + 2*x+rnorm(10,sd=10)
```

The likelihood ratio test

```
plot(x,y)
```



The likelihood ratio test

```
## null hypothesis model:
m0<-lm(y~1)
## alternative hypothesis model:
m1<-lm(y~x)
```

```
deltalambda<- -2*(logLik(m0)-logLik(m1))
## observed value:
deltalambda[1]

## [1] 9.584

## critical value:
qchisq(0.95,df=1)

## [1] 3.8415
```

The likelihood ratio test

```
# p-value:
pchisq(deltalambda[1],df=1,lower.tail=FALSE)

## [1] 0.0019628
```

The likelihood ratio test

Here, we fit the null hypothesis model which only has an intercept term β_0, and the alternative model that has β_1 as well.
Finally, we compare the $\Delta\Lambda$ with the critical chi-squared value for degrees of freedom 1.
We also computed the probability of getting a $\Delta\Lambda$ as extreme as we got assuming that the null is true:

The likelihood ratio test

Note that in the likelihood test above, we are comparing one nested model against another: the null hypothesis model is nested inside the alternative hypothesis model.
What this means is that the alternative hypothesis model contains all the parameters in the null hypothesis model (i.e., the intercept) plus another one (the slope).

ANOVA

We can compare two models, one nested inside another, as follows:

```
anova(m0,m1)

## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1      9 468
## 2      8 179  1       288 12.9 0.0071
```

ANOVA

The F-score you get here is actually the square of the t-value you get in the linear model summary:

```
sqrt(anova(m0,m1)$F[2])

## [1] 3.5861

summary(m1)$coefficients[2,3]

## [1] 3.5861
```

This is because $t^2 = F$. The proof is discussed on page 9 of the Dobson and Barnett book.

ANOVA

The ANOVA works as follows. First define the residual as:[1]

$$e = Y - X\hat{\beta} \tag{39}$$

The square of this is:

$$e^T e = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \tag{40}$$

[1] Note that I do not use ε here, but e, to refer to the residual. This is because these are the estimates of ε derived from the estimates $\hat{\beta}$.

ANOVA

Define the **deviance** as:

$$\begin{aligned} D =& \frac{1}{\sigma^2}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\ =& \frac{1}{\sigma^2}(Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta}) \\ =& \frac{1}{\sigma^2}(Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}) \end{aligned} \tag{41}$$

ANOVA

Linear modeling theory tells us $\hat{\beta} = (X^T X)^{-1} X^T Y$.
Premultiplying both sides with $(X^T X)$, we get

$$(X^T X)\hat{\beta} = X^T Y$$

It follows that we can rewrite the last line in equation 41 as follows: We can replace $(X^T X)\hat{\beta}$ with $X^T Y$.

$$\begin{aligned} D =& \frac{1}{\sigma^2}(Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T \underline{X^T X\hat{\beta}}) \\ =& \frac{1}{\sigma^2}(Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T \underline{X^T Y}) \\ =& \frac{1}{\sigma^2}(Y^T Y - Y^T X\hat{\beta}) \end{aligned} \tag{42}$$

ANOVA

Notice that $Y^T X \hat{\beta}$ is a scalar (1×1) and is identical to $\beta^T X^T Y$ (check this), so we could write:

ANOVA

$$D = \frac{1}{\sigma^2}(Y^T Y - \hat{\beta}^T X^T Y)$$

Assume now that we have data of size n.
Suppose we have a null hypothesis $H_0 : \beta = \beta_0$ and an alternative hypothesis $H_1 : \beta = \beta_1$.
Let the null hypothesis have q parameters, and the alternative p, where $q < p < n$.
Let X_0 be the design matrix for H_0, and X_1 the design matrix for H_1.

ANOVA

Compute the deviances D_0 and D_1 for each hypothesis, and compute ΔD:

$$\begin{aligned}\Delta D =& D_0 - D_1 = \frac{1}{\sigma^2}[(Y^T Y - \hat{\beta}_0 X_0^T Y) - (Y^T Y - \hat{\beta}_1 X_1^T Y)] \\ =& \frac{1}{\sigma^2}[\hat{\beta}_1 X_1^T Y - \hat{\beta}_0 X_0^T Y]\end{aligned} \tag{43}$$

ANOVA

It turns out that the F-statistic has the following distribution if the null hypothesis is true:

$$F = \frac{\Delta D/(p-q)}{D_1/(n-p)} \sim F(p-q, n-p) \tag{44}$$

So, an extreme value of F is inconsistent with the null and we reject it.

ANOVA

The F-statistic is:

$$
\begin{aligned}
F =& \frac{\Delta D/(p-q)}{D_1/(n-p)} \\
=& \frac{\hat{\beta}_1 X_1^T Y - \hat{\beta}_0^T X_0^T Y}{p-q} \Big/ \frac{Y^T Y - \hat{\beta}_1^T X_1^T Y}{n-p}
\end{aligned} \tag{45}
$$

ANOVA

Traditionally, the way the F-test is summarized is:

Table: ANOVA table

Source of variance	df	Sum of squares	Mean square
Model with β_0	q	$\beta_0^T X_0^T Y$	
Improvement due to β_1	p-q	$\hat{\beta}_1 X_1^T Y - \hat{\beta}_0^T X_0^T Y$	$\frac{\hat{\beta}_1 X_1^T Y - \hat{\beta}_0^T X_0^T Y}{p-q}$
Residual	n-p	$Y^T Y - \hat{\beta}_1^T X_1^T Y$	$\frac{Y^T Y - \hat{\beta}_1^T X_1^T Y}{n-p}$
Total	n	$y^T y$	

There is much more to say here about ANOVA, but this is the basic idea.

ANOVA

Practically speaking, the usage is simple:

```
anova(m0,m1)

## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1      9 468
## 2      8 179  1       288 12.9 0.0071
```

Here, the F-statistic tells us that the model m1 is "better" (there is a significant effect of the predictor x).
Whether that is a meaningful result depends on the power properties of the design (to be discussed later).