

Introduction to statistics: Linear models

Shravan Vasishth

Universität Potsdam
vasishth@uni-potsdam.de
<http://www.ling.uni-potsdam.de/~vasishth>

April 4, 2020

Summary

1. We learnt about the single sample, two sample, and paired t-tests.
2. We learnt about Type I, II error (and power).
3. We learnt about Type M and Type S errors.

Now we are ready to look at linear modeling.

Load Grodner and Gibson dataset

```
gge1crit<-read.table("data/grodnergibson05data.txt",header=1)  
head(gge1crit)
```

##	subject	item	condition	rawRT
## 6	1	1	objgap	320
## 19	1	2	subjgap	424
## 34	1	3	objgap	309
## 49	1	4	subjgap	274
## 68	1	5	objgap	333
## 80	1	6	subjgap	266

Compute means by factor level

Let's compute the means by factor levels:

```
means<-round(with(gge1crit,tapply(rawRT,IND=condition,  
                                   mean)))
```

```
means
```

```
##   objgap subjgap
```

```
##      471      369
```

The object relative mean is higher than the subject relative mean.

Paired t-test on the data

Correct t-test by subject and by items

This is how one would do a t-test CORRECTLY with such data, to compare means across conditions:

```
bysubj<-aggregate(rawRT~subject+condition,
                  mean,data=ggelcrit)
byitem<-aggregate(rawRT~item+condition,mean,data=ggelcrit)
t.test(rawRT~condition,paired=TRUE,bysubj)$statistic

##          t
## 3.1093

t.test(rawRT~condition,paired=TRUE,byitem)$statistic

##          t
## 3.7542
```

Paired t-test on the data

Consider only by-subject analyses for now.

These are the means we are comparing by subject:

```
round(with(bysubj,  
tapply(rawRT,condition,mean)))
```

```
##  objgap subjgap  
##    471    369
```

Linear models

We can rewrite our best guess about how the object and subject relative clause reading time distributions like this:

Object relative: $Normal(471, \hat{\sigma})$

Subject relative: $Normal(471 - 102, \hat{\sigma})$

Note that the two distributions for object and subject relative are assumed to be independent. This is not true in our data as we get a data point each for each RC type from the same subject!

Linear models

- ▶ The object relative's distribution can be written as a sum of two terms:

$$y = 471 + \epsilon \text{ where } \epsilon \sim \text{Normal}(0, \hat{\sigma})$$

- ▶ The subject relative's distribution can be written:

$$y = 471 - 102 + \epsilon \text{ where } \epsilon \sim \text{Normal}(0, \hat{\sigma})$$

- ▶ Note that $\hat{\sigma} = 213$ because $obs.t = \frac{\bar{x}}{s/\sqrt{n}} \Rightarrow s =$

$$\bar{x} \times \sqrt{n}/obs.t = -103 \times \sqrt{42}/-3.109 = 213.$$

The above statements describe a generative process for the data.

Linear models

Now consider this **linear model**, which describes the rt in each row of the data frame as a function of condition. ϵ is a random variable $\epsilon \sim Normal(0, 213)$.

Object relative reading times:

$$rt = 471 + \epsilon \quad (1)$$

Subject relative reading times:

$$rt = 471 - 102 + \epsilon \quad (2)$$

Linear models

When describing mean reading times, I can drop the ϵ :
Object relative reading times:

$$rt = 471 \quad (3)$$

Subject relative reading times:

$$rt = 471 - 102 \quad (4)$$

The `lm()` function gives us these mean estimates from the data.

Linear models

Object relative reading times:

$$rt = 471 \times \mathbf{1} - 102 \times \mathbf{0} + \epsilon \quad (5)$$

Subject relative reading times:

$$rt = 471 \times \mathbf{1} - 102 \times \mathbf{1} + \epsilon \quad (6)$$

So, object relatives are coded as 0, and subject relatives are coded as 1.

The `lm()` function sets up such a model.

Linear models

With real data from the relative clause study:

```
contrasts(bysubj$condition)

##           subjgap
## objgap           0
## subjgap           1

m0<-lm(rawRT~condition,bysubj)
round(summary(m0)$coefficients)[,1]

##      (Intercept) conditionsubjgap
##              471             -102
```

Linear models

The linear model gives us two numbers: object relative reading time (471), and the difference between object and subject relative (-102):

```
round(coef(m0))
```

```
##      (Intercept) conditionsubjgap  
##              471             -102
```

Linear models

1. The **intercept** is giving us the mean of the objgap condition.
2. The **slope** is giving us the amount by which the subject relative is faster.

Note that the meaning of the intercept and slope depends on the ordering of the factor levels. We can make subject relative means be the intercept:

```
## reverse the factor level ordering:
bysubj$condition<-factor(bysubj$condition,
                          levels=c("subjgap", "objgap"))
contrasts(bysubj$condition)

##           objgap
## subjgap         0
## objgap          1
```

Linear models

```
m1a<-lm(rawRT~condition,bysubj)
round(coef(m1a))

##      (Intercept) conditionobjgap
##             369             102
```

Now the intercept is the subject relative clause mean.
The slope is the increase in reading time for the object relative condition.

Linear models

```
## switching back to the original
## factor level ordering:
bysubj$condition<-factor(bysubj$condition,
                          levels=c("objgap", "subjgap"))
contrasts(bysubj$condition)

##           subjgap
## objgap           0
## subjgap           1
```


Linear models

In mathematical form, the model is:

$$rt = \beta_0 + \beta_1 condition + \epsilon \quad (7)$$

where

- ▶ β_0 is the mean for the object relative
- ▶ β_1 is the amount by which the object relative mean must be changed to obtain the mean for the subject relative.

The null hypothesis is that the difference in means between the two relative clause types β_1 is:

$$H_0 : \beta_1 = 0$$

Linear models

The **contrast coding** determines the meaning of the β parameters:

```
bysubj$condition<-factor(bysubj$condition,  
                           levels=c("objgap","subjgap"))  
contrasts(bysubj$condition)
```

```
##           subjgap  
## objgap           0  
## subjgap           1
```

Linear models

We will make a distinction between the **unknown true mean** β_0, β_1 and the **estimated mean from the data** $\hat{\beta}_0, \hat{\beta}_1$.

- ▶ Estimated mean object relative processing time: $\hat{\beta}_0 = 471$.
- ▶ Estimated mean subject relative processing time:
 $\hat{\beta}_0 + \hat{\beta}_1 = 471 + -102 = 369$.

Linear models

Reparameterizing the linear model with sum contrast coding

In mathematical form, the model is:

$$rt = \beta_0 + \beta_1 condition + \epsilon \quad (8)$$

We can change the **contrast coding** to change the meaning of the β parameters:

```
## new contrast coding:  
bysubj$cond<-ifelse(bysubj$condition=="objgap",1,-1)
```

Linear models

Reparameterizing the linear model with sum contrast coding

```
xtabs(~cond+condition,bysubj)
```

```
##      condition
## cond objgap subjgap
##   -1      0      42
##    1     42      0
```

Linear models

Reparameterizing the linear model with sum contrast coding

Now the model parameters have a different meaning:

```
m1<-lm(rawRT~cond,bysubj)
round(coef(m1))
```

## (Intercept)	cond
## 420	51

Linear models

Reparameterizing the linear model with sum contrast coding

- ▶ Estimated **grand mean** processing time: $\hat{\beta}_0 = 420$.
- ▶ Estimated mean object relative processing time:
 $\hat{\beta}_0 + \hat{\beta}_1 = 420 + 51 = 471$.
- ▶ Estimated mean subject relative processing time:
 $\hat{\beta}_0 - \hat{\beta}_1 = 420 - 51 = 369$.

This kind of parameterization is called **sum-to-zero contrast** or more simply **sum contrast** coding. This is the coding we will use.

Linear models

The null hypothesis for the slope

The null hypothesis for the slope is

$$H_0 : \mathbf{1} \times \mu_{obj} + (-\mathbf{1} \times) \mu_{subj} = 0 \quad (9)$$

The sum contrasts are referring to the ± 1 terms in the null hypothesis:

- ▶ object relative: +1
- ▶ subject relative: -1

Linear models

Now the model is:

Object relative reading times:

$$rt = 420 \times \mathbf{1} + 51 \times \mathbf{1} + \epsilon \quad (10)$$

Subject relative reading times:

$$rt = 420 \times \mathbf{1} + 51 \times -\mathbf{1} + \epsilon \quad (11)$$

So, object relatives are coded as 1, and subject relatives are coded as -1.

Linear models

The model is:

$$rt = \beta_0 + \beta_1 + \epsilon \text{ where } \epsilon \sim \text{Normal}(0, \sigma) \quad (12)$$

It is an assumption of the linear model that the residuals are (approximately) normally distributed.

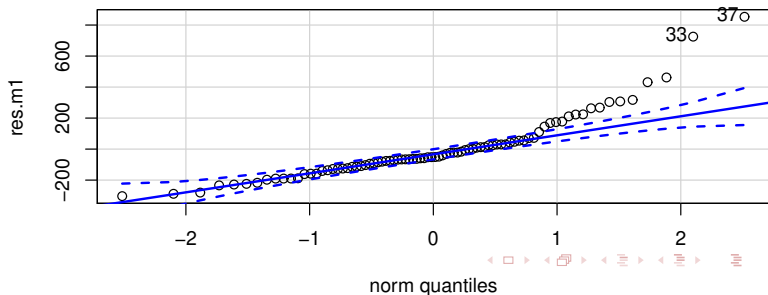
We can check that this assumption is met:

```
## residuals:  
res.m1<-residuals(m1)
```

Linear models

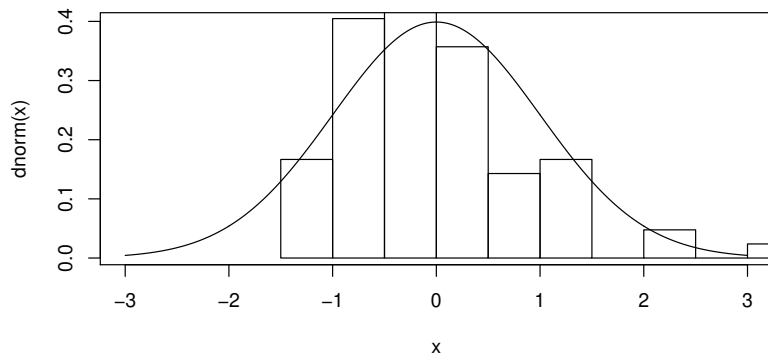
Plot the residuals by comparing them to the standard normal distribution (Normal(0,1)):

```
## [1] 37 33
```



Linear models

Another way to visualize it (not the standard way):



Linear models

Log transformation

A log-transform improves the normality of residuals:

```
m1log<-lm(log(rawRT)~cond,bysubj)
round(coef(m1log),4)
```

```
## (Intercept)      cond
##      5.9488      0.0843
```

Linear models

Log transformation

- ▶ Estimated **grand mean** processing time: $\hat{\beta}_0 = 5.9488$.
- ▶ Estimated mean object relative processing time:
 $\hat{\beta}_0 + \hat{\beta}_1 = 5.9488 + 0.0843 = 6.0331$.
- ▶ Estimated mean subject relative processing time:
 $\hat{\beta}_0 - \hat{\beta}_1 = 5.9488 - 0.0843 = 5.8645$.

Linear models

The model is:

$$\log rt = \beta_0 + \beta_1 + \epsilon \quad (13)$$

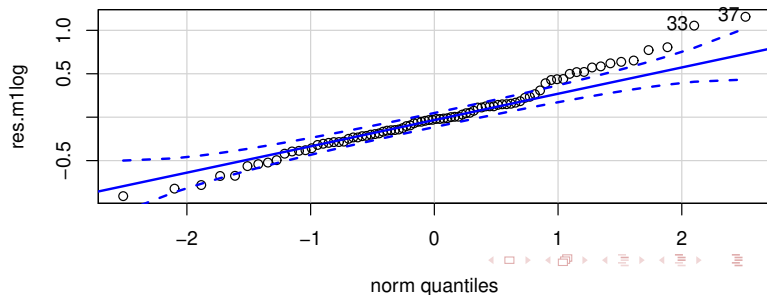
Now check the residuals:

```
## residuals:  
res.m1log<-residuals(m1log)
```

Linear models

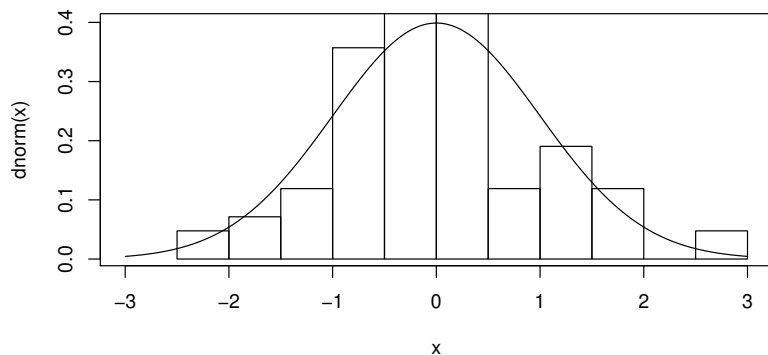
Plot the residuals by comparing them to the standard normal distribution (Normal(0,1)):

```
## [1] 37 33
```



Linear models

Another way to visualize it (not the standard way):



Linear models

Log transformation: recovering estimates on ms scale

- ▶ Estimated mean object relative processing time:
 $\hat{\beta}_0 + \hat{\beta}_1 = 5.9488 + 0.0843 = 6.0331.$
- ▶ Estimated mean subject relative processing time:
 $\hat{\beta}_0 - \hat{\beta}_1 = 5.9488 - 0.0843 = 5.8645.$

Note that $\exp(\log(rt)) = rt.$

To get the mean estimates on the raw ms scale, we just exponentiate both sides of the equation:

$$\exp(\log rt) = \exp(\beta_0 + \beta_1)$$

Linear models

Log transformation: recovering estimates on ms scale

- ▶ Estimated mean object relative processing time:
 $\exp(\hat{\beta}_0 + \hat{\beta}_1) = \exp(5.9488 + 0.0843) = 417.$
- ▶ Estimated mean subject relative processing time:
 $\exp(\hat{\beta}_0 - \hat{\beta}_1) = \exp(5.9488 - 0.0843) = 352.$

The difference in reading time is $417 - 352 = 65$ ms (cf. 102 ms on raw scale).

Linear models: Summary

Summary

Using the relative clause example, we learnt

- ▶ the meaning of treatment contrast coding.
- ▶ the meaning of sum contrast coding. **This is the coding we will use.**

For a more comprehensive discussion of contrast coding, read this paper:

How to capitalize on a priori contrasts in linear (mixed) models: A tutorial, Schad, Hohenstein, Vasishth, Kliegl. Download from:

<https://arxiv.org/abs/1807.10451>

Paired t-tests vs linear models

The linear mixed model

The paired t-test and the linear model t-test values don't match:

```
t.test(rawRT~condition,bysubj,paired=TRUE)$statistic
```

```
##          t
```

```
## 3.1093
```

```
round(summary(m0)$coefficients,2)[,c(1:3)]
```

```
##              Estimate Std. Error t value
```

```
## (Intercept)      471.36      31.13   15.14
```

```
## conditionsubjgap -102.29      44.02   -2.32
```

Paired t-tests vs linear models

The linear mixed model

This is because the linear model implements the **unpaired** (i.e., two sample) t-test:

```
round(summary(m0)$coefficients,2)[,c(1:3)]
```

##	Estimate	Std. Error	t value
## (Intercept)	471.36	31.13	15.14
## conditionsubjgap	-102.29	44.02	-2.32

```
round(t.test(rawRT~condition,bysubj,  
             paired=FALSE)$statistic,2)
```

##	t
##	2.32

Paired t-tests vs linear models

The linear mixed model

The paired t-test has an equivalent in the linear modeling framework:

```
m0.lmer<-lmer(rawRT~condition+(1|subject),bysubj)
summary(m0.lmer)$coefficients
```

##	Estimate	Std. Error	t value
## (Intercept)	471.36	31.128	15.1428
## conditionsubjgap	-102.29	32.896	-3.1093

We turn to the linear mixed model in the next lecture.

Linear models

1. In our relative clause example, the 'predictor' is categorical.
2. What about when we have continuous predictors?
3. For example, we have instructors' "beauty" levels measured on a continuous scale as predictors of their teaching evaluations.
4. Beauty levels are centered; this means that a beauty level of 0 means average beauty level. This is a data set from a paper by Hamermesh and Parker (Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity," Economics of Education Review, August 2005). I got the data from Gelman and Hill (2007).

Linear models

```
bdata <- read.table("data/beauty.txt",header=T)
head(bdata)
```

```
##      beauty evaluation
## 1  0.20157          4.3
## 2 -0.82608          4.5
## 3 -0.66033          3.7
## 4 -0.76631          4.3
## 5  1.42145          4.4
## 6  0.50022          4.2
```

```
plot(evaluation~beauty,bdata)
```

Linear models

Note that the beauty scores are centered to have mean (approximately) 0:

```
summary(bdata$beauty)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	-1.5388	-0.7446	-0.1564	-0.0883	0.4572	1.8817

Linear models

One model we can fit:

$$y = \beta_0 + \epsilon$$

```
m2<-lm(evaluation~1,bdata)
mean(bdata$evaluation)

## [1] 3.9983

round(summary(m2)$coefficients,digits=3)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.998      0.026   155.05      0
```

This model is only estimating the grand mean of evaluation scores.

Linear models

$$y = \beta_0 + \beta_1 x + \epsilon$$

```
m3<-lm(evaluation~beauty,bdata)
round(summary(m3)$coefficients,digits=3)
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	4.010	0.026	157.205	0
## beauty	0.133	0.032	4.133	0

The intercept now means: the expected evaluation score given an average beauty level.

The slope means: the expected increase in evaluation with **one unit** increase in beauty.

Summary

We now know how to fit

1. Simple linear models with a categorical predictor (relative clause data)
2. Simple linear models with a continuous predictor.

The likelihood ratio test

Sometimes, students ask me about ANOVA. I generally avoid discussing the ANOVA, but if anyone is interested, here are the details.

This section is optional. I encourage you to work through it, but understanding this material is not necessary for the rest of the course.

The likelihood ratio test

Define the likelihood ratio statistic as:

$$\Lambda = \frac{\max_{\theta \in \omega_0}(\text{lik}(\theta))}{\max_{\theta \in \omega_1}(\text{lik}(\theta))} \quad (14)$$

where, $\omega_0 = \{\mu_0\}$ and $\omega_1 = \{\forall \mu \mid \mu \neq \mu_0\}$.

Suppose that X_1, \dots, 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:

$$\frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2\right) \quad (15)$$

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

$$\frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2\right) \quad (16)$$

The likelihood ratio test

The likelihood ratio statistic is then:

$$\Lambda = \frac{\frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2\right)}{\frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2\right)} \quad (17)$$

The likelihood ratio test

Canceling out common terms:

$$\Lambda = \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2\right)}{\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2\right)} \quad (18)$$

The likelihood ratio test

Taking logs:

$$\begin{aligned}\log \Lambda &= \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2 \right) - \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \right) \\ &= -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \bar{X})^2 \right)\end{aligned}\quad (19)$$

The likelihood ratio test

Now, note that

$$\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 \quad (20)$$

The likelihood ratio test

This means that

$$\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \bar{X})^2 = n(\bar{X} - \mu_0)^2 \quad (21)$$

The likelihood ratio test

So, we can write Λ as:

$$\Lambda = -\frac{1}{2\sigma^2}n(\bar{X} - \mu_0)^2 \quad (22)$$

Rearranging terms:

$$-2\Lambda = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} \quad (23)$$

The likelihood ratio test

Or maybe even more transparently:

$$-2\Lambda = \frac{(\bar{X} - \mu_0)^2}{\frac{\sigma^2}{n}} \quad (24)$$

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

The likelihood ratio test

More generally, we will define the **likelihood ratio test statistic** as follows. Here, $L(\theta)$ refers to the likelihood given some value of θ , and $\ell(\theta)$ refers to the log likelihood.

$$\begin{aligned} L &= -2 \times \log(L(\theta_0)/L(\theta_1)) \\ \log L &= -2 \times \{\ell(\theta_0) - \ell(\theta_1)\} \\ \log L &= -2 \times \{\ell(\theta_0) - \ell(\theta_1)\} \end{aligned} \tag{25}$$

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 $\log L$ is sufficiently large. As the sample size approaches infinity:

$$\log L \rightarrow \chi_r^2 \text{ as } n \rightarrow \infty \tag{26}$$

where r is called the degrees of freedom and is the difference in the number of parameters under H_1 and H_0 .

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:

$$\log L = 2\{\ell(\theta_1) - \ell(\theta_0)\} \quad (27)$$

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.

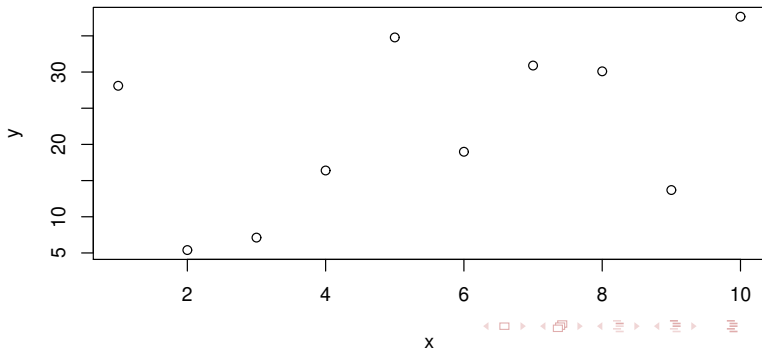
The likelihood ratio test

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)
```

```
lambda<- -2*(logLik(m0)-logLik(m1))  
## observed value:  
lambda[1]  
  
## [1] 2.3342  
  
## critical value:  
qchisq(0.95,df=1)  
  
## [1] 3.8415
```

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 λ with the critical chi-squared value for degrees of freedom 1.

We also computed the probability of getting a λ 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 plus some others.

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 1196
## 2          8  947  1      249 2.1   0.19
```


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] 1.4503  
  
summary(m1)$coefficients[2,3]  
## [1] 1.4503
```

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

$$e = Y - X\hat{\beta} \quad (28)$$

The square of this is:

$$e^T e = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \quad (29)$$

¹Note 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:

$$\begin{aligned} D &= \frac{1}{\sigma^2} (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \\ &= \frac{1}{\sigma^2} (Y^T - \hat{\beta}^T X^T) (Y - X\hat{\beta}) \\ &= \frac{1}{\sigma^2} (Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}) \end{aligned} \tag{30}$$

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 30 as follows: We can replace $(X^T X)\hat{\beta}$ with $X^T Y$.

$$\begin{aligned} D &= \frac{1}{\sigma^2} (Y^T Y - Y^T X \hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T \underline{X^T X} \hat{\beta}) \\ &= \frac{1}{\sigma^2} (Y^T Y - Y^T X \hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T \underline{X^T Y}) \\ &= \frac{1}{\sigma^2} (Y^T Y - Y^T X \hat{\beta}) \end{aligned} \tag{31}$$

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 :

$$\begin{aligned}\Delta D = D_0 - D_1 &= \frac{1}{\sigma^2} [(Y^T Y - \hat{\beta}_0 X_0^T Y) - (Y^T Y - \hat{\beta}_1 X_1^T Y)] \\ &= \frac{1}{\sigma^2} [\hat{\beta}_1 X_1^T Y - \hat{\beta}_0 X_0^T Y]\end{aligned}\tag{32}$$

ANOVA

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

$$F = \frac{\Delta D / (p - q)}{D_1 / (n - p)} \sim F(p - q, n - p) \quad (33)$$

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

ANOVA

The F-statistic is:

$$\begin{aligned} F &= \frac{\Delta D / (p - q)}{D_1 / (n - p)} \\ &= \frac{\hat{\beta}_1 X_1^T Y - \hat{\beta}_0^T X_0^T Y}{p - q} / \frac{Y^T Y - \hat{\beta}_1^T X_1^T Y}{n - p} \end{aligned} \quad (34)$$

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 1196
## 2      8  947  1      249 2.1  0.19
```

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).