

Statistical Methods for Research II

Shizhe Chen

2020-02-03

Contents

Preface	5
1 Causal Inference	7
1.1 Association and Causality	7
1.2 Potential Outcomes	7
1.3 Experiments v.s. Observational Studies	7
1.4 Learning Objectives	7
2 One-way ANOVA	9
2.1 Simple randomized experiments	9
2.2 One-way ANOVA	9
2.3 Model diagnostics	12
2.4 Learning Objectives	18
3 Two-way ANOVA	21
3.1 Experiments with two (or more) factors	21
3.2 Two-way ANOVA	25
3.3 Learning Objectives	36
4 Random and Mixed Effect Models	39
4.1 Nested design	39
4.2 Random effects model	39
4.3 Learning Objectives	39
5 Repeated Measures Design	41
5.1 Repeated measures design	41
5.2 Analysis of repeated measures designs	41
6 Case-control Study	45
6.1 Case-control study	45
6.2 Logistic regression	45
6.3 Generalized linear model	45
7 Observational Study	47
7.1 Causality in observational studies	47

7.2	Analysis with no latent confounding	47
7.3	Instrumental variable	47
7.4	Missing data	48
8	Complex data	49
8.1	Data in the big data era	49
8.2	Useful methods	49
9	Project description	51
9.1	Project 1: Project STAR I	51
9.2	Project 2: Project STAR II	52
9.3	Background	52
9.4	Tasks	52
9.5	Project 3: US Traffic Fatalities	53
9.6	Background	53
9.7	Tasks	53
9.8	Hints	54
9.9	Project 4: Bank Marketing	54
9.10	Background	54
9.11	Tasks	54
9.12	Hints	55

Preface

This file contains code and comments in STA 207. Many materials in this document are adapted from lectures by Professors Prabir Burman, Peng Ding, Lukas Meier, Kosuke Imai, and Zhichao Jiang.

Chapter 1

Causal Inference

1.1 Association and Causality

1.2 Potential Outcomes

- Definition of potential outcomes
- Properties of potential outcomes
- Key assumptions in this notation
- Immutable characteristics

1.3 Experiments v.s. Observational Studies

- Advantages of randomized experiments
- Reasons for observational studies

1.4 Learning Objectives

- Students are able to distinguish association and causality.
- Students gain basic familiarity with the potential outcome framework.
- Students are able to recognize the importance of experiment designs.

Chapter 2

One-way ANOVA

2.1 Simple randomized experiments

- Motivation and real world applications of randomized experiments.
- Sampling schemes of a simple randomized experiment.
- Question of interest, null hypotheses, and their causal interpretation.
- Intuition of hypothesis testing.
- Practical concerns in the design of experiments.

2.2 One-way ANOVA

2.2.1 A motivating example: the Spock trial

In 1968 Dr. Benjamin Spock was tried in Boston for conspiring against the government for helping young men to escape the military draft. He was convicted by the Boston federal court, but the judgement was overturned by the Court of Appeals in 1969 for many reasons, one of which was cited as the bias of the presiding judge Francis Ford. Dr. Spock, a pediatrician, was very famous for his books on rearing of children, and thus was widely admired by women. As a matter of fact, the jury in Spock trial has no women. Note that jury panels, though randomly selected, should reflect the demographics. In any particular trial, there may not be any woman on the jury, but it is worthwhile to examine if the jury panels of Judge Ford had fewer women than other judges in Boston in few months before the trial. Data are available for jury panels for 7, but we investigate the data for only 4 judges including Judge Ford.

```
Spock <- read.csv(file="./data/SpockTrial.csv", header=TRUE, sep=",")
Spock$Judge<-as.factor(Spock$Judge);
# Box plot with jittered points (from stackoverflow: https://stackoverflow.com/question)
boxplot(perc.women~Judge,data=Spock)
stripchart(perc.women~Judge, vertical = TRUE, data = Spock,
  method = "jitter", add = TRUE, pch = 20, col = 'blue')
```

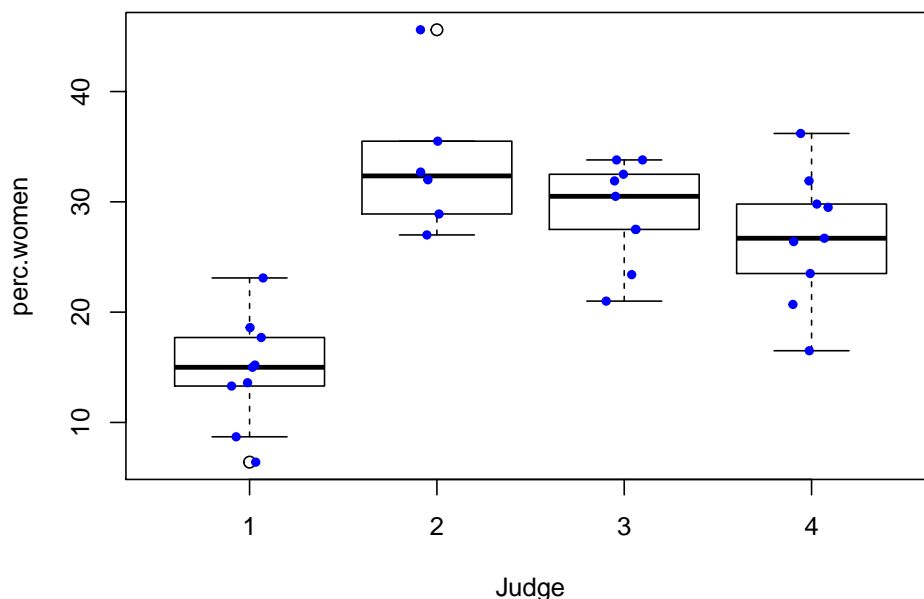


Figure 2.1: Box plot with jittered data points for the Spock trial data.

2.2.2 ANOVA model

- Cell means model
- Estimators of the means
- Decomposition of sum of squares
- Some basic properties

In the Spock trial data, we can use `aov()` to fit a one-way ANOVA model.

```
anova.fit<- aov(perc.women~Judge,data=Spock)
summary(anova.fit)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Judge      3   1591     530    17.6 1.1e-06 ***
## Residuals  29    874      30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can obtain the following.

- $n_1 = 9$, $n_2 = 6$, $n_3 = 9$, $n_4 = 9$, and $n_T = n_1 + n_2 + n_3 + n_4 = 33$.
- $\bar{Y}_{1\cdot} = 14.62$, $\bar{Y}_{2\cdot} = 33.62$, $\bar{Y}_{3\cdot} = 29.1$, and $\bar{Y}_{4\cdot} = 26.8$.
- $SSE = 873.52$, $df = 29$, and $MSE = 30.12$.
- $SSTO = 2464.8$, $df = 32$.
- $SSTR = 1591.28$, $df = 3$, and $MSTR = 530.43$.

2.2.3 Statistical inference

- Null hypothesis
- The F-test
- Testing a linear combination
 - Estimation
 - Hypothesis testing
- (Simultaneous) confidence intervals

To test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against the alternative $H_1 : \text{not all } \mu_i\text{'s are equal}$. We can calculate the F-statistics $F^* = \frac{MSTR}{MSE} = 17.61$, when $F(0.95; 3, 29) = 2.93$. We can thus reject the null hypothesis at the chosen significance level 0.05.

Consider the quantity $L = \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3$. We can calculate that $\hat{L} = -15.22$, and $s^2(\hat{L}) = 4.65$. Moreover, since $t(1 - 0.01/2; n_T - r) = 2.76$ a 99% confidence interval for L is $(-21.16, -9.27)$.

To test the hypothesis $H_0 : L = 0$ against $H_1 : L \neq 0$. We can calculate the t-statistics $t^* = \hat{L}/s(\hat{L}) = -7.06$. We can calculate the quantile of the t-distribution as before to finish the test.

Next we demonstrate how to construct the simultaneous confidence intervals for all pairwise difference. The code will construct confidence intervals for two pairs, but leave you to finish the rest.

```
# Create vectors for the first two linear combinations
comb.mat<-matrix(0,nrow=2,ncol=4)
comb.mat[1,]=c(1,-1,0,0);comb.mat[2,]=c(1,0,-1,0);

# Obtain the estimates
diff = numeric(dim(comb.mat)[1]);
diff.sd=diff;
mean.tmp=anova.fit$coefficients;mean.tmp[1]=0;
ns=as.numeric(table(Spock$Judge));
for(i in 1:length(diff)){
  diff[i]=sum(comb.mat[i,]*mean.tmp);
  diff.sd[i]=sqrt(sum(comb.mat[i,]^2/ns)*mse);
}

alpha=0.05;

# Bonferroni correction:
m=6; # for all pairwise differences, although we only show two here
B.stat=qt(1-alpha/(2*m),anova.fit$df.residual);

# Tukey-Kramer
T.stat=qtukey(1-alpha, nmeans=length(anova.fit$coefficients), df=anova.fit$df.residual),
```

```

# Scheffe
S.stat=sqrt( (length(anova.fit$coefficients)-1)*qf(1-alpha,length(anova.fit$coefficients)

table.stats=matrix(0,1,3);
table.stats[1,]=c(B.stat,T.stat,S.stat);
colnames(table.stats)=c('Bonferroni', 'Tukey', 'Scheffe')
table.stats

##      Bonferroni Tukey Scheffe
## [1,]          2.8   2.7      3

# Then, we can construct the confidence intervals as, e.g.,
CI.bonferroni =matrix(0,nrow=2,ncol=2);
for(i in 1:length(diff)){
  CI.bonferroni[i,]=diff[i]+c(1,-1)*B.stat*diff.sd[i];
}

```

2.2.4 Alternative forms of the ANOVA model

- Factor-effect model
- Regression model

The default of ANOVA in R set weights proportional to the sample size in each cell. You can supply the `weights` to the `aov()` function to force equal weights.

```

# Weights proportional to sample sizes
print(model.tables(anova.fit,"effects"))

```

```

## Tables of effects
##
## Judge
##      1      2      3      4
## -10.72 8.271 3.755 1.455
## rep   9.00 6.000 9.000 9.000

```

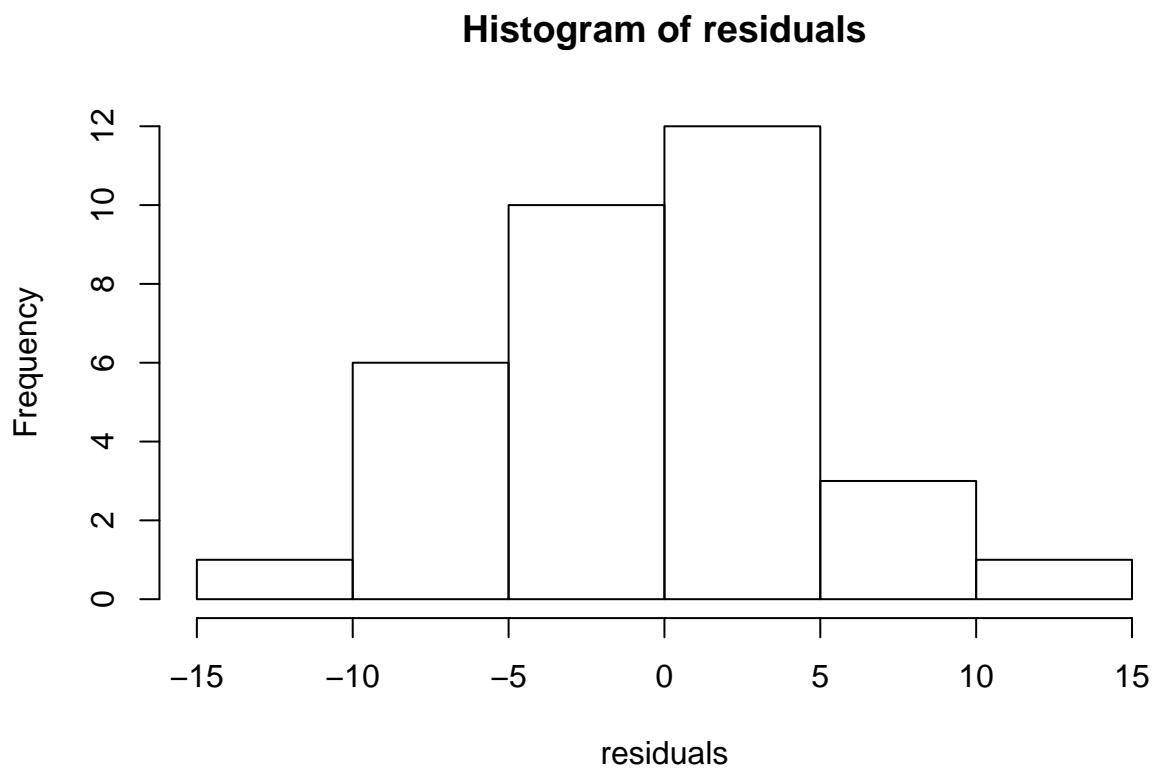
2.3 Model diagnostics

- Possible departures
- Diagnostics with residuals
- Graphical methods
- Formal tests
 - Hartley test
 - Bartlett test
 - Levene test
- Remedial measures

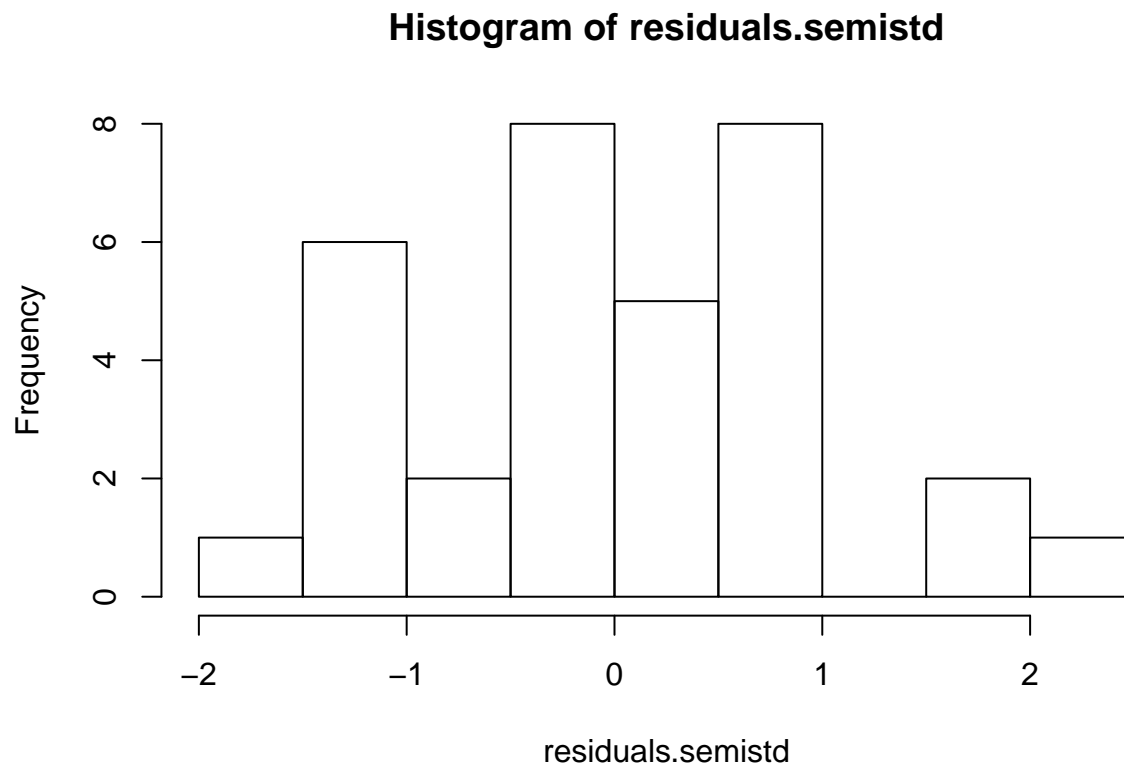
- Weighted least squares
- Nonparametric tests based on ranks: rank test, Kruskal-Wallis test
- Box-Cox transformation

All diagnostics start with the residuals.

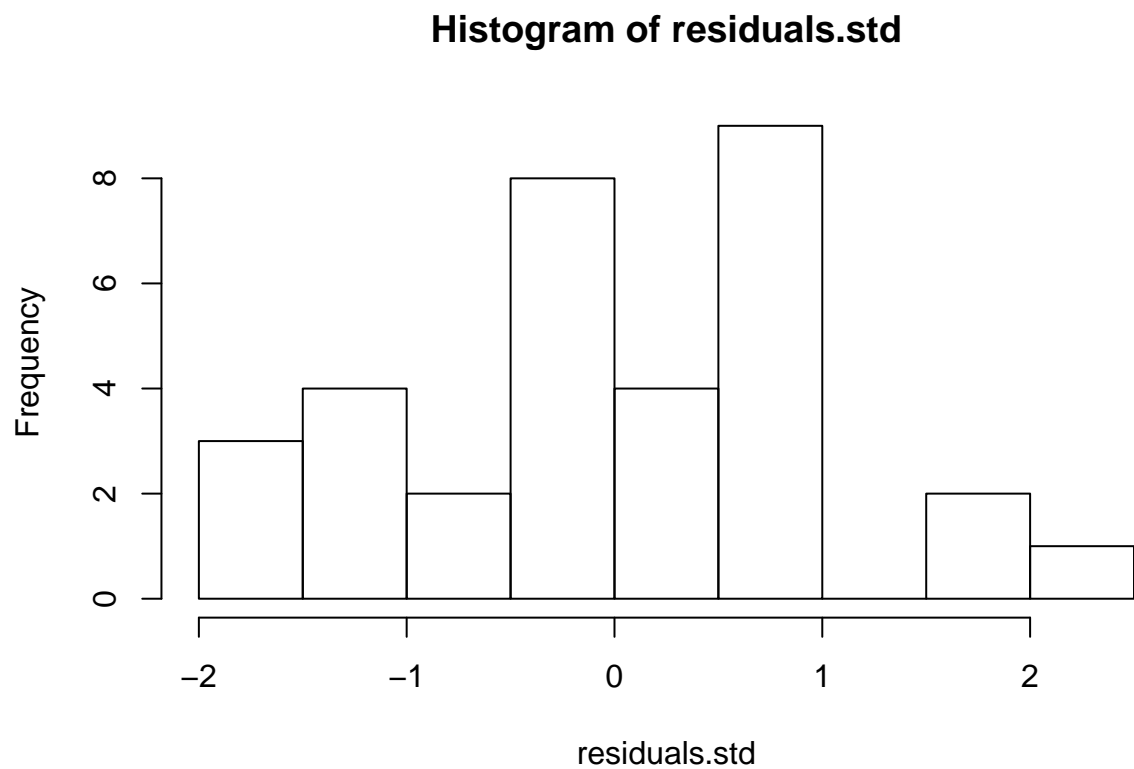
```
# Obtain the residuals from the ANOVA fit  
residuals=anova.fit$residuals;  
hist(residuals)
```



```
# Semistudentized residuals  
residuals.semistd=anova.fit$residuals/sqrt(mse);  
hist(residuals.semistd)
```

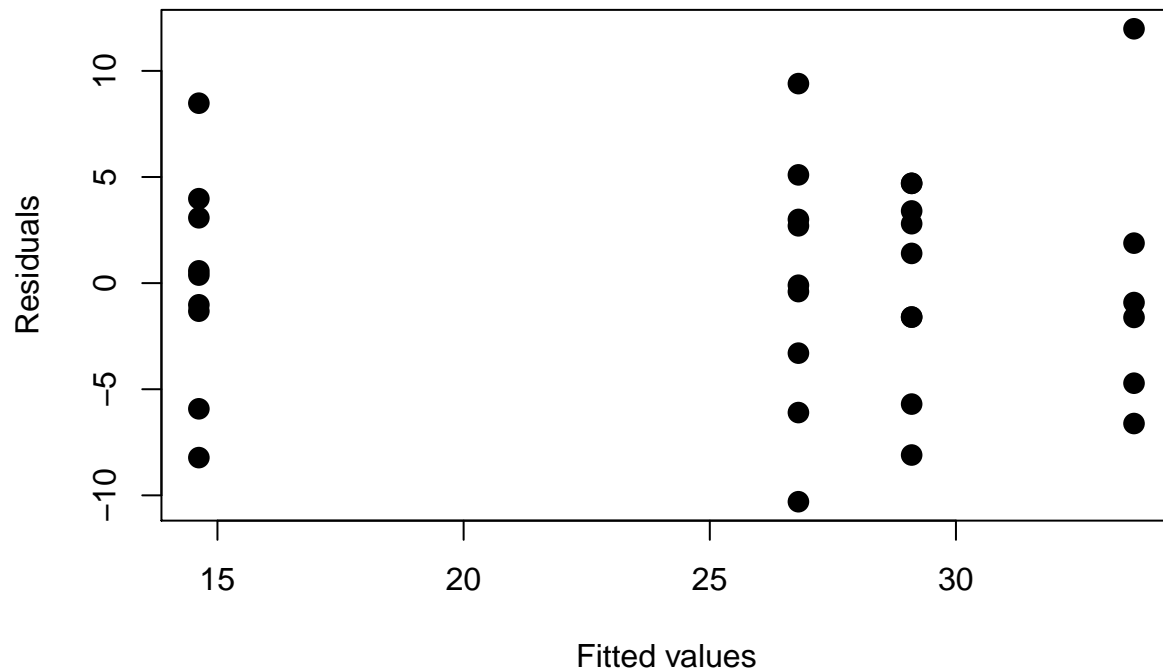


```
# Studentized residuals
weights=1-1/ns[as.numeric(Spock$Judge)];
residuals.std=anova.fit$residuals/sqrt(mse)/sqrt(weights);
hist(residuals.std)
```



Plot the residuals (or the other two versions) against fitted values

```
plot(residuals~anova.fit$fitted.values,type='p',pch=16,cex=1.5,xlab="Fitted values",ylab="Residuals")
```



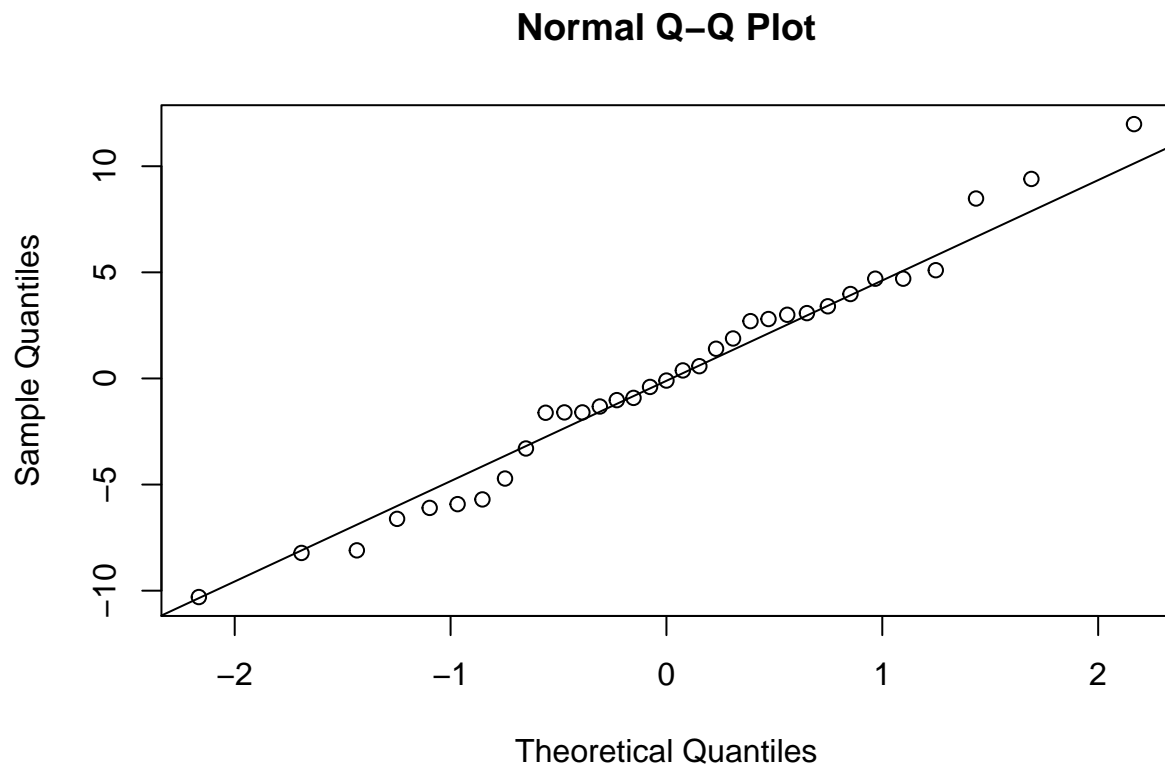
```
# Plot the residual against certain orders
# No clear orders make sense in the Spock trial data
```

```
# Stem-leaf plot (or use histogram, or qq-plot )
stem(residuals)
```

```
##
## The decimal point is at the |
##
## -10 | 3
## -8 | 21
## -6 | 61
## -4 | 977
## -2 | 3
## -0 | 66630941
## 0 | 4649
## 2 | 78014
## 4 | 0771
## 6 |
## 8 | 54
## 10 |
## 12 | 0
```



```
qqnorm(residuals);qqline(residuals)
```



```
# Plot residuals against missing variables  
# Not applicable on Spock trial data
```

We now turn to formal tests of the equality of variances.

```
# Calculate the variances for each group:  
(vars = tapply(Spock$perc.women, Spock$Judge, var))
```

```
##  1  2  3  4  
## 25 43 21 36
```

```
alpha=0.05;
```

```
# Hartley test:  
H.stat=max(vars)/min(vars);  
library(SuppDists) # The distribution is in this package  
# Both df and k only take integers:  
qmaxFratio(1-alpha, df=floor(sum(ns)/length(ns)-1), k=length(ns))
```

```
## [1] 8.4
```

```
qmaxFratio(1-alpha,df=ceiling(sum(ns)/length(ns)-1),k=length(ns))
```

```
## [1] 7.2
```

```
# Bartlett test:
```

```
K.stat= (sum(ns)-length(ns))*log(mse)-sum( (ns-1)*log(vars) );
qchisq(1-alpha,df=length(ns)-1)
```

```
## [1] 7.8
```

```
# Levene test:
```

```
Spock$res.abs=abs(anova.fit$residuals);
summary(aov(res.abs~Judge,data=Spock))
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## Judge      3    5.6    1.88    0.17  0.91
## Residuals  29  314.7   10.85
```

We leave weighted least squares for exercise. You can either calculate it following the steps discussed in lecture, or use the `weights` option in `lm()` and `aov()`.

We can conduct the nonparametric tests as follows.

```
# The rank test
```

```
Spock$rank.perc=rank(Spock$perc.women)
summary(aov(rank.perc~Judge,data=Spock))
```

```
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Judge      3   1846     615    15.6 3.1e-06 ***
## Residuals  29   1144      39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Kruskal-Wallis test:
```

```
kruskal.test(perc.women~Judge,data=Spock)
```

```
##
```

```
##  Kruskal-Wallis rank sum test
```

```
##
```

```
## data:  perc.women by Judge
```

```
## Kruskal-Wallis chi-squared = 20, df = 3, p-value = 2e-04
```

For Box-Cox transformation, use the `boxcox` in library `MASS`.

2.4 Learning Objectives

- Students are able to write down a one-way ANOVA model given a new dataset.
- Students understand the basic properties of one-way ANOVA models.

- Students recognize the assumptions associated with each method.
- Students can implement the aforementioned tasks in R.
- Students are comfortable reading R helpfiles related to one-way ANOVA.

Chapter 3

Two-way ANOVA

3.1 Experiments with two (or more) factors

“Block what you can and randomize what you cannot.”

— Box et al. (2005)

- Randomized experiments with two treatments
- Stratified randomized Experiments (also known as randomized block design)
 - Auditor training data
 - Project STAR
- Reasons for stratification: practical and statistical
- Sampling scheme for a stratified randomized experiment
- Question of interest, null hypotheses, and their causal interpretation.
- Intuition of hypothesis testing.

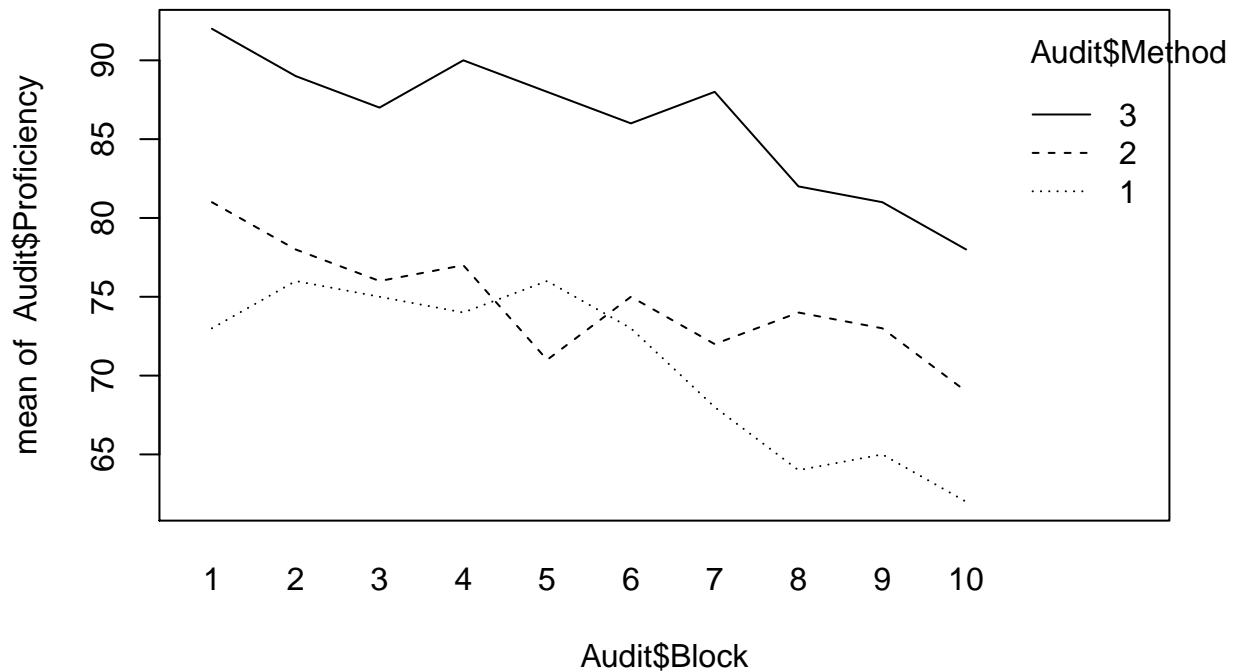
Auditor training

We can check whether stratification is efficient on the auditor training data.

Description of the auditor training data: There are three training methods for the auditors and the response Y is a proficiency score after the training are completed. Ideally we would like to compare the three methods among those who are as similar as possible in their educational background. How do we achieve this? One way is compare the three different training methods among those whose time since graduation from college are about the same. Suppose then we have ten such groups (of three individuals each). Group 1 consists of those who graduated recently, group 2 people graduated between one and two years ago, and group 10 consists of those who graduated some time in the past (say, ten years or more). Time since graduation is called the block (or a blocking factor) and treatment is the training method.

```
Audit <- read.csv(file="./data/AuditorTraining.csv", header=TRUE, sep=",")

# Draw the interaction plot:
interaction.plot(Audit$Block, Audit$Method, Audit$Proficiency)
```



```
# Fit the model with blocks:
anova.block<-aov(Proficiency~as.factor(Block)+as.factor(Method),data=Audit)
anova.random<-aov(Proficiency~as.factor(Method),data=Audit)

mse.block<-sum(anova.block$residuals^2)/anova.block$df.residual
mse.random<-sum(anova.random$residuals^2)/anova.random$df.residual
(E=mse.random/mse.block)
```

```
## [1] 3.2
```

Simulation

We can also explore the properties of blocking using simulation.

```
# A simple simulation for stratification
set.seed(10928)
# Data generating mechanism:
n=40;n.strata=10;
X= sample(x=(1:n.strata),size=n,replace=TRUE);
ACE=4; coef.X=2;
Y.1=ACE+coef.X*X+rnorm(n,mean=0,sd=1); # potential outcome
Y.0=coef.X*X+rnorm(n,mean=0,sd=1); # potential outcome
trt= sample(1:n,size=(n/2),replace=FALSE);Z=rep(0,n);Z[trt]=1; # randomization
```

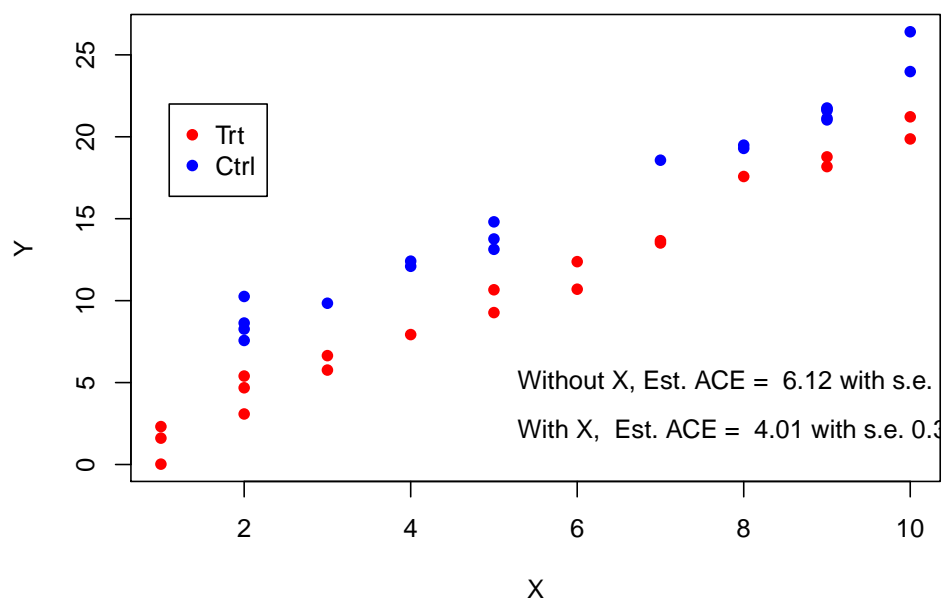


Figure 3.1: Estimation of ACE with and without adjusting for X

```
Z.s=rep(0,n);
for (i in 1:n.strata){# randomization within stratum
  id.stratum= which(X==i);
  trt= sample(id.stratum,size=floor(length(id.stratum)/2),replace=FALSE);
  Z.s[trt]=1;
}
```

```
Y=Y.1*Z+Y.0*(1-Z); # observation w/o stratification
Y.s=Y.1*Z.s+Y.0*(1-Z.s); # observation w stratification
```

```
# Analysis, w/o stratification
```

```
lm.vanilla=summary(lm(Y~Z));
lm.X=summary(lm(Y~Z+X));
```

```
plot(y=Y,x=X,pch=16,col=c('red','blue')[Z+1])
legend(x=1.1,y=22,legend=c('Trt','Ctrl'),col=c('red','blue'),pch=16)
text(x=5.1,y=5,labels=paste('Without X, Est. ACE = ',round(lm.vanilla$coef[2,1],2), 'with s.e. ',round(lm.vanilla$coef[2,2],2)))
text(x=5.1,y=2,labels=paste('With X, Est. ACE = ',round(lm.X$coef[2,1],2), 'with s.e. ',round(lm.X$coef[2,2],2)))
```

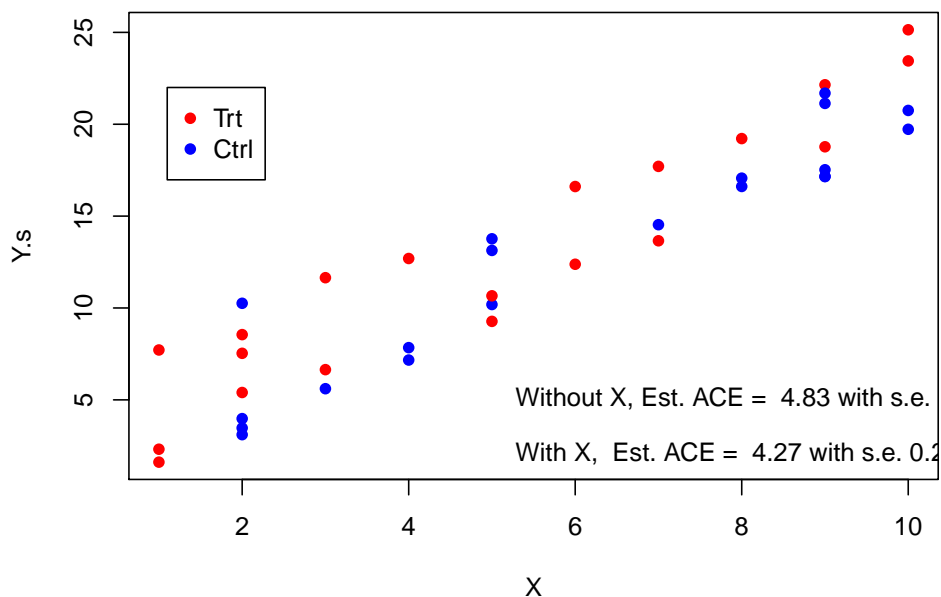


Figure 3.2: Estimation of ACE with and without adjusting for X under stratified randomization

```
# Analysis, w/o stratification
```

```
lm.simple=summary(lm(Y.s~Z.s));
```

```
lm.strat=summary(lm(Y.s~Z.s+X));
```

```
plot(y=Y.s,x=X,pch=16,col=c('red','blue')[Z+1])
```

```
legend(x=1.1,y=22,legend=c('Trt','Ctrl'),col=c('red','blue'),pch=16)
```

```
text(x=5.1,y=5,labels=paste('Without X, Est. ACE = ',round(lm.simple$coef[2,1],2), 'with s.e. ',round(lm.simple$se[2,1],2)))
```

```
text(x=5.1,y=2,labels=paste('With X, Est. ACE = ',round(lm.strat$coef[2,1],2), 'with s.e. ',round(lm.strat$se[2,1],2)))
```

```
# Repeat the above procedure 10000 times to evaluate the efficiency
```

```
# Wrap up the code in one function
```

```
strat.sim<-function(ACE){
```

```
n=40;n.strata=10;
```

```
X= sample(x=(1:n.strata),size=n,replace=TRUE);
```

```
coef.X=5;
```

```
Y.1=ACE+coef.X*X+rnorm(n,mean=0,sd=1); # potential outcome
```

```
Y.0=coef.X*X+rnorm(n,mean=0,sd=1); # potential outcome
```



```

trt= sample(1:n,size=(n/2),replace=FALSE);Z=rep(0,n);Z[trt]=1; # randomization

Z.s=rep(0,n);
for (i in 1:n.strata){# randomization within stratum
  id.stratum= which(X==i);
  trt= sample(id.stratum,size=floor(length(id.stratum)/2),replace=FALSE);
  Z.s[trt]=1;
}

Y=Y.1*Z+Y.0*(1-Z); # observation w/o stratification
Y.s=Y.1*Z.s+Y.0*(1-Z.s); # observation w stratification

lm.vanilla=summary(lm(Y~Z));
lm.X=summary(lm(Y~Z+X));

lm.simple=summary(lm(Y.s~Z.s));
lm.strat=summary(lm(Y.s~Z.s+X));

est.ACE=c(lm.vanilla$coef[2,1],lm.X$coef[2,1],lm.simple$coef[2,1],lm.strat$coef[2,1])
return(est.ACE)
}
ACE=4;

sim.result=replicate(n=1e4,strat.sim(ACE=ACE));

(mse=apply(sim.result-ACE,MARGIN=1,sd))

## [1] 4.59 0.32 1.36 0.32

```

Q: Why conducting stratify randomization given the simulation results?

For more discussion, see Chapters 6 and 8 in ANOVA: A Short Intro Using R by Lukas Meier.

3.2 Two-way ANOVA

3.2.1 A motivating example: Hey fever relief data set

For the Hay Fever Relief example, 9 compounds for Hay Fever Relief are made by varying levels of the two basic ingredients. Ingredient 1 (factor A) has $a = 3$ levels: low ($i = 1$), medium ($i = 2$) and high ($i = 3$). Similarly, ingredient 2 (factor B) has $b = 3$ levels: low ($j = 1$), medium ($j = 2$) and high ($j = 3$). A total of 36 subjects (suffering from hay fever) are selected and each of the 9 compounds are given to randomly selected $n = 4$ individuals.

3.2.2 A two-way ANOVA model

- Cell mean model
- Decomposition of the means, and their estimators
- Additive models
 - Why additive models?
 - Estimators of the means
- Decomposition of sum of squares, and their properties

3.2.3 Statistical inference

- F-statistics based on sums of squares
- Hypothesis testing
 - Test for interaction effects
 - Test for main effects
 - Alternative test if interaction can be ignored (additive models)
- (Simultaneous) confidence intervals with and without interactions
 - Bonferroni
 - Tukey
 - Scheffe

3.2.4 Model diagnostics

- Similar to those for one-way ANOVA

3.2.5 Strategy for data analysis

Using the Hey Fever data as an example.

Step 0. Exploratory data analysis.

```
Hay <- read.csv(file="./data/HayFever.csv", header=TRUE, sep=",")
```

```
# Use a slightly different visualization:
```

```
pairs(Hay,pch=16,col='red',cex=1.5)
```

```
# Or draw the main effect plots
```

```
plotmeans(Relief~Ingredient.1,data=Hay,xlab="Ingredient 1",ylab="Relief", main="Main ef
```

```
plotmeans(Relief~Ingredient.2,data=Hay,xlab="Ingredient 2",ylab="Relief", main="Main ef
```

```
interaction.plot(Hay$Ingredient.1, Hay$Ingredient.2, Hay$Relief)
```

Step 1. Test whether interaction effects are presented.

```
# We can use the regression form here
```

```
full_model=lm(Relief~as.factor(Ingredient.1)+as.factor(Ingredient.2)+as.factor(Ingredient
```

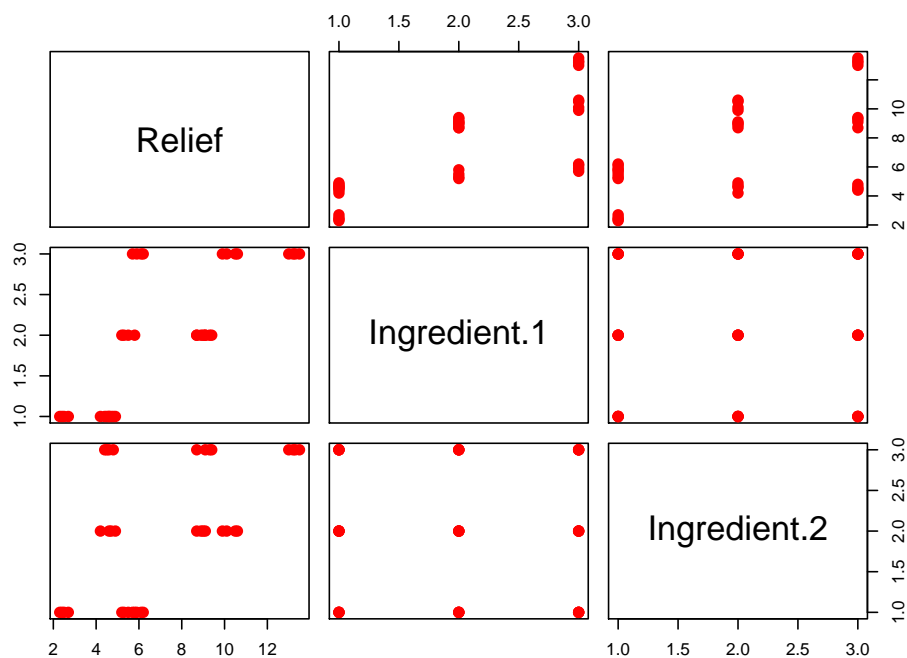


Figure 3.3: Box plot with jittered data points for the Hay Fever data.

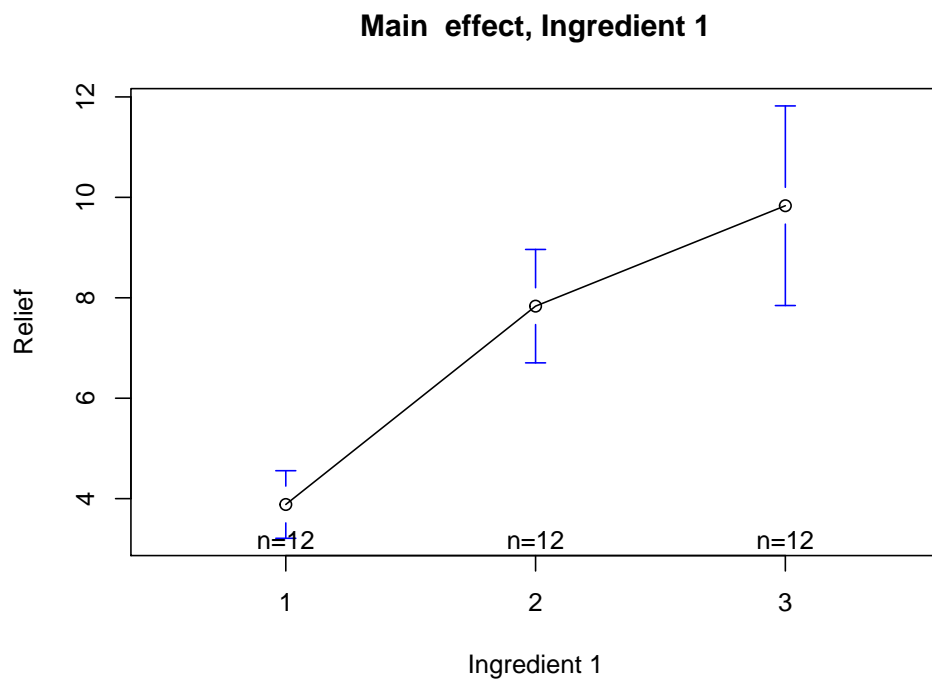


Figure 3.4: Main effect plot for ingredient 1

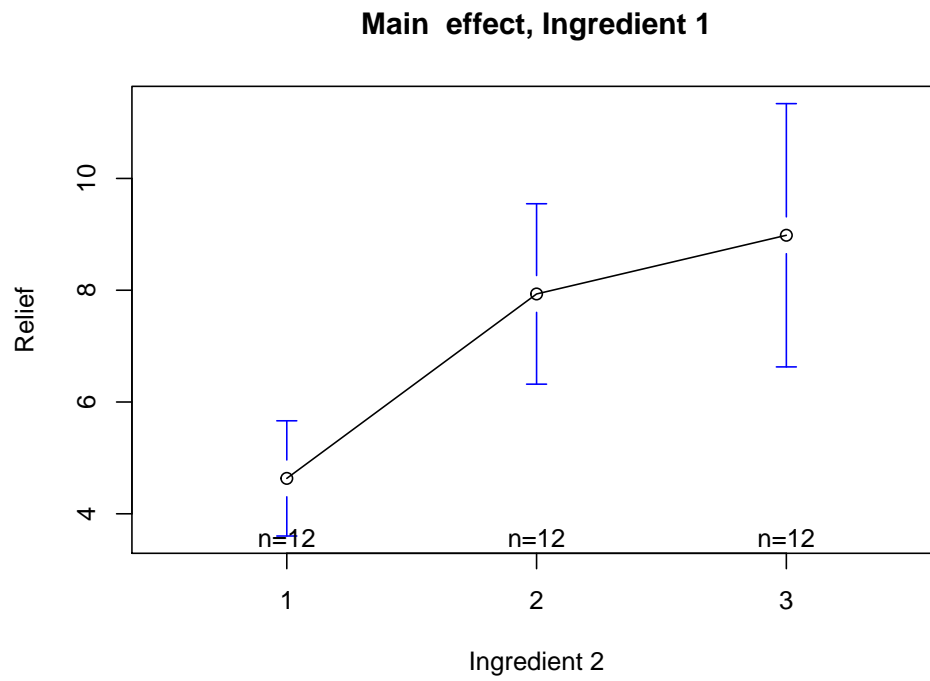


Figure 3.5: Main effect plot for ingredient 2

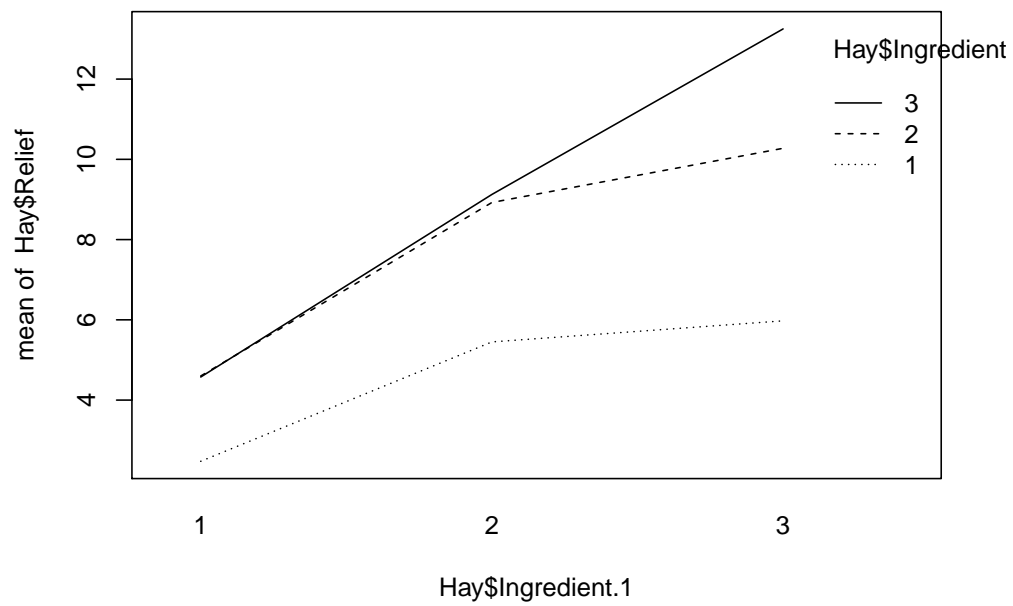


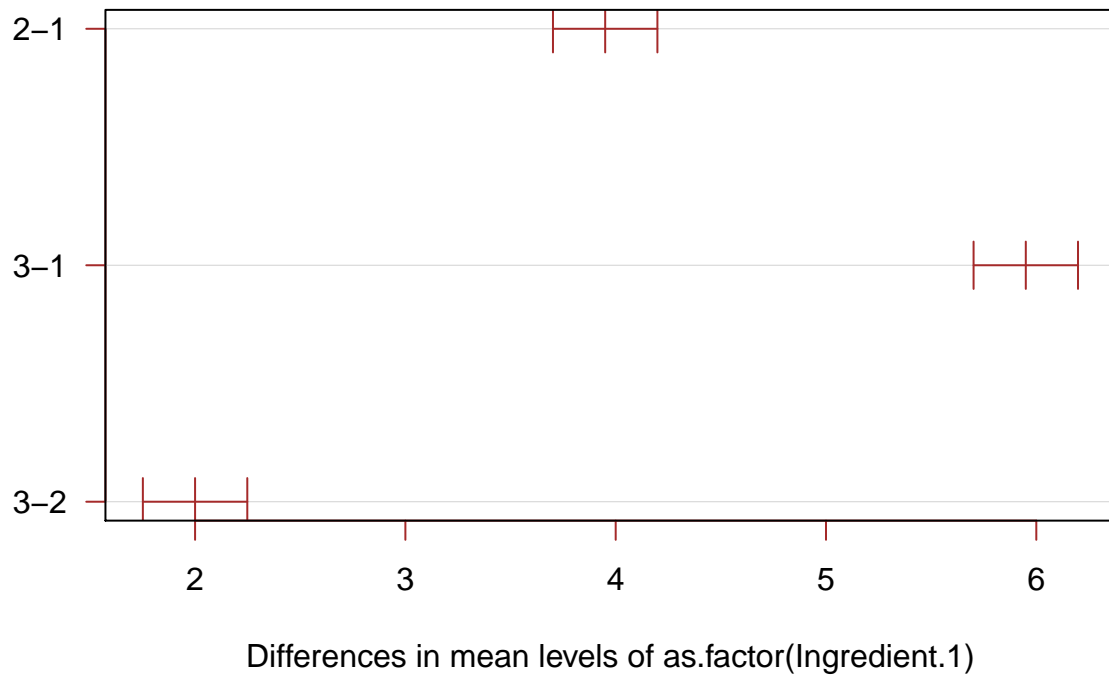
Figure 3.6: Interaction plot

```
reduced_model=lm(Relief~as.factor(Ingredient.1)+as.factor(Ingredient.2),data=Hay);
anova(reduced_model,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: Relief ~ as.factor(Ingredient.1) + as.factor(Ingredient.2)
## Model 2: Relief ~ as.factor(Ingredient.1) + as.factor(Ingredient.2) +
##      as.factor(Ingredient.1) * as.factor(Ingredient.2)
##   Res.Df    RSS Df Sum of Sq   F Pr(>F)
## 1      31 31.05
## 2      27  1.63  4      29.4 122 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

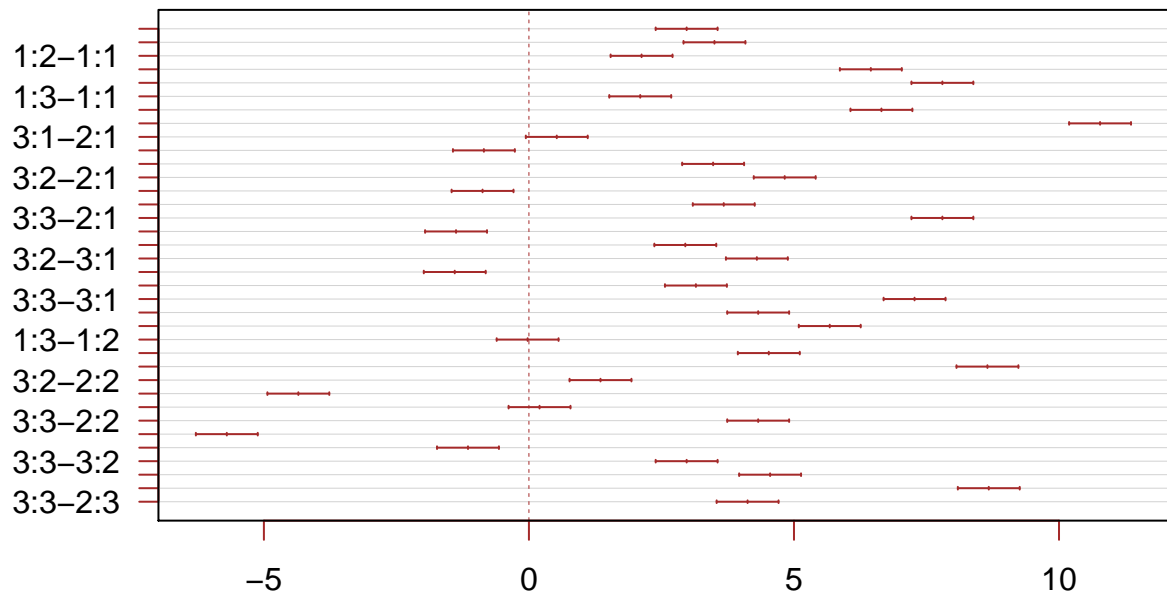
The test result show that interaction effects are very unlikely to be absent from this data set. This means that we need to treat each combination as a unit, whereas we can compare each type of main effects separately. In the Hay Fever data, we naturally want to find the combination of ingredients that is most effective. We can use the Tukey-Kramer method for this task.

```
library(stats)
alpha=0.05;
anova.fit<-aov(Relief~as.factor(Ingredient.1)+as.factor(Ingredient.2)+as.factor(Ingredient.3))
T.ci=TukeyHSD(anova.fit,conf.level = 1-alpha)
plot(T.ci, las=1 , col="brown")
```

95% family-wise confidence level



95% family-wise confidence level

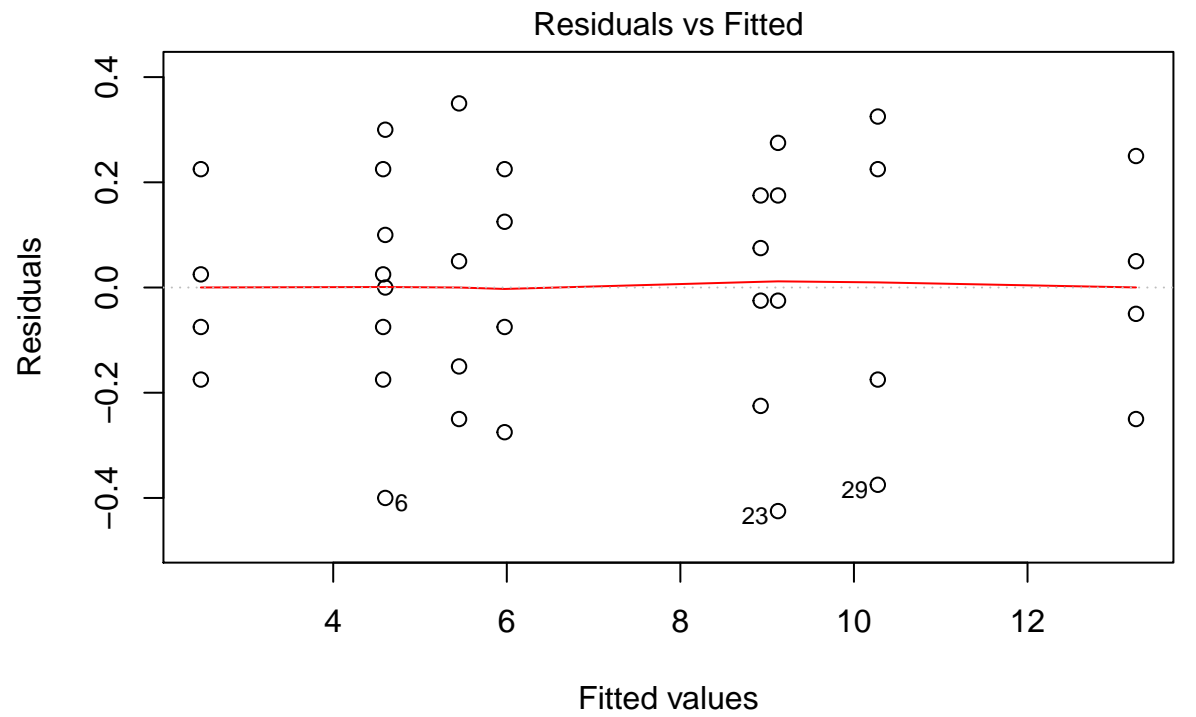


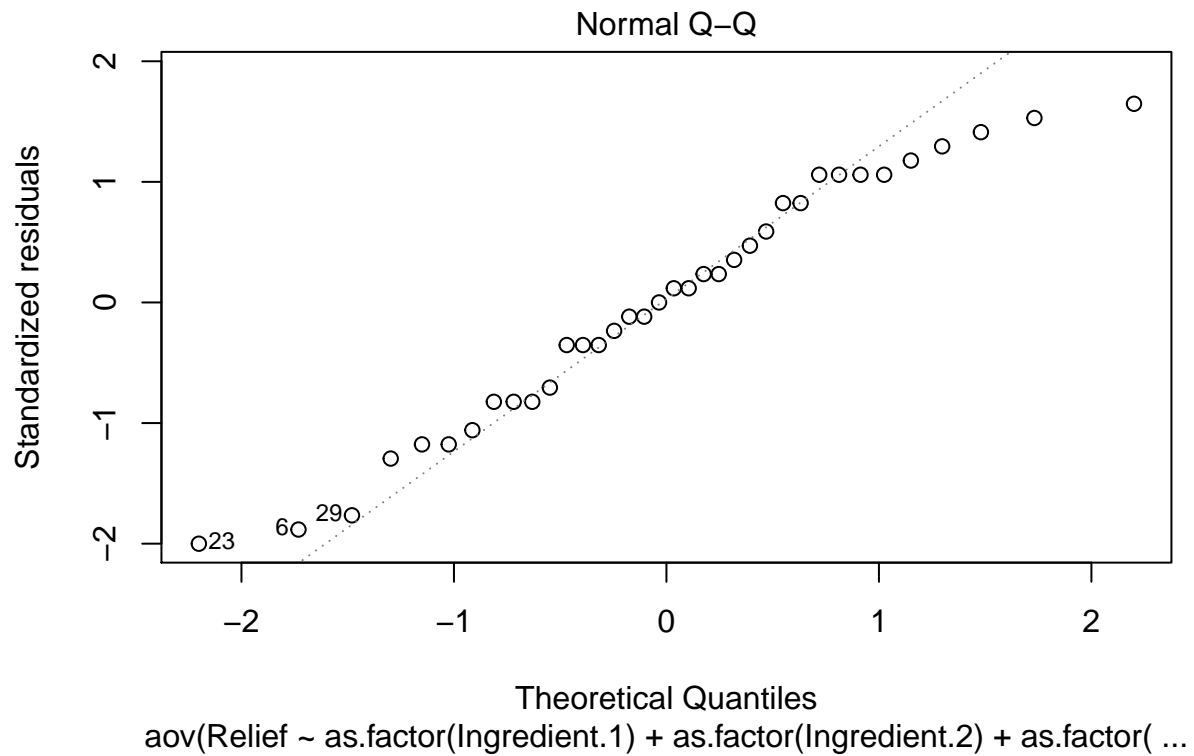
Differences in mean levels of as.factor(Ingredient.1):as.factor(Ingredient.2)

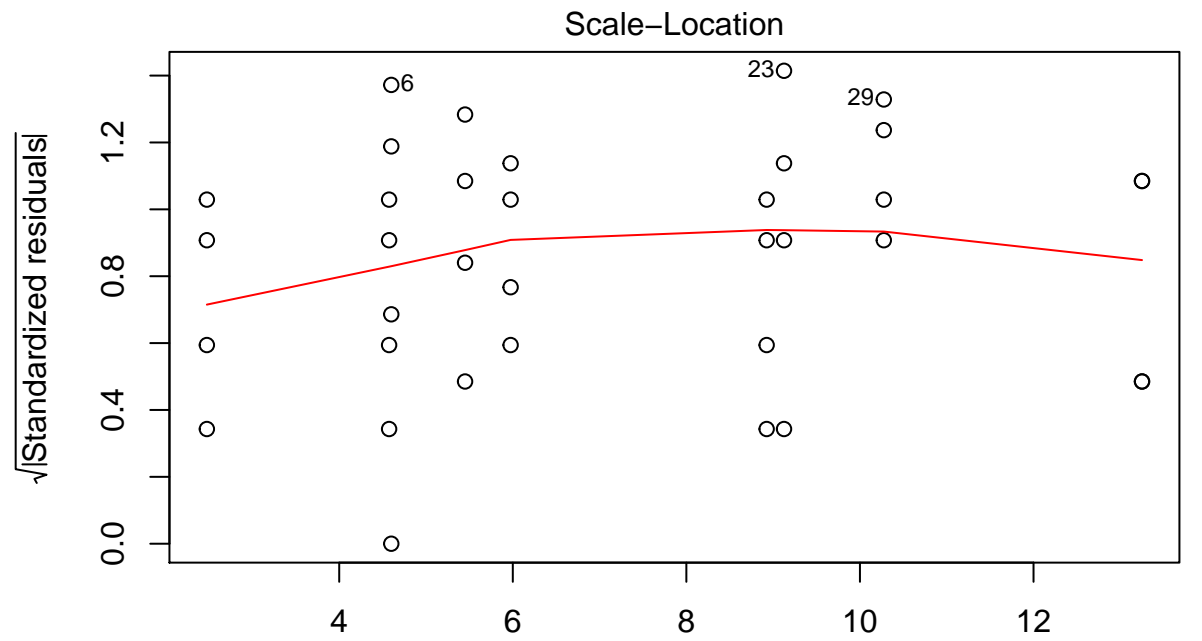
```
# We only need to pay attention to the differences of the two largest means
idx=list();
idx[[1]]=Hay$Ingredient.1;idx[[2]]=Hay$Ingredient.2;
(means.comb=tapply( Hay$Relief, INDEX=idx,mean))
```

```
##      1      2      3
## 1 2.5  4.6  4.6
## 2 5.4  8.9  9.1
## 3 6.0 10.3 13.2
```

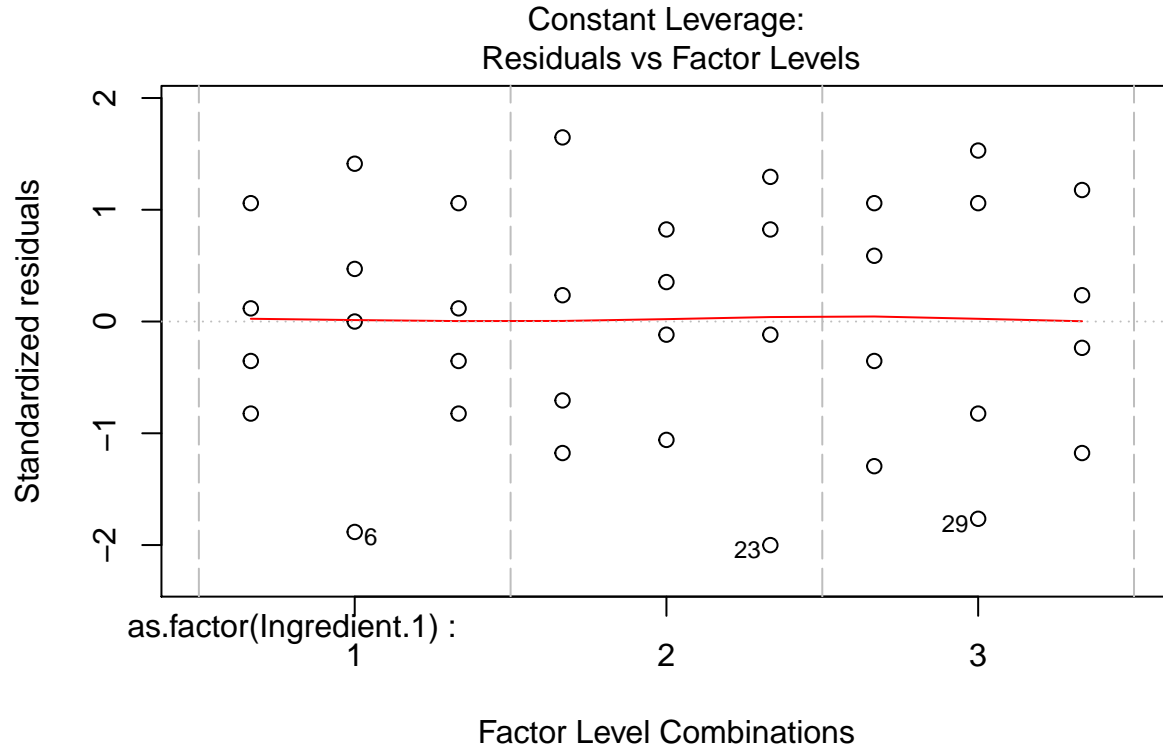
```
# For model diagnostics, we can use the default plotting function of aov()
plot(anova.fit)
```





Fitted values
aov(Relief ~ as.factor(Ingredient.1) + as.factor(Ingredient.2) + as.factor(...



3.2.6 Special case: one observation per cell

- Interaction effects are no longer identifiable
- Estimation and testing
- Tukey's test of additivity

3.2.7 Unbalanced two-way ANOVA

- The model is the same as for the balanced case
- Estimators for the means and variances
- Hypothesis testing
 - Interaction effects
 - Main effects
- Missing data in the one observation per cell case

3.3 Learning Objectives

- Students are able to write down an appropriate two-way ANOVA model given a new dataset.
- Students understand the basic properties of the various types of two-way ANOVA models.

- Students recognize the assumptions associated with each method, and can find appropriate tests to verify the assumptions.
- Students can implement the aforementioned tasks in R.
- Students can seek help in coding using the Internet.

Chapter 4

Random and Mixed Effect Models

(In progress)

4.1 Nested design

- Motivations for a nested design
- Nested design with fixed factors
 - Sampling scheme
 - Hypothesis testing
 - Causal interpretation
- Nested design with random factors
- Nested design with mixed factors
 - Repeated measures design

4.2 Random effects model

- One-way ANOVA model with random effects
- Estimation
 - Decomposition of variances
- Hypothesis testing and confidence intervals
- Mixed effects model
- Unbalanced mixed and random effect models

For more discussion, see Chapter 7 here.

4.3 Learning Objectives

- Students are able to write down the two-way ANOVA model with random effects.
- Students can properly decide whether to model a factor using fixed or random effects.
- Students recognize the key assumptions associated with the random effects model.

- Students can implement the aforementioned tasks in R.
- Students can explore extension of random effect model on the Internet.

Chapter 5

Repeated Measures Design

(In progress)

5.1 Repeated measures design

- Motivation for repeated measures
- Sampling scheme
- Estimation, hypothesis, and causal interpretation
- Split plot design
- Longitudinal data
 - Experiments
 - Observational studies: prospective and retrospective cohort study
 - Sampling scheme for observational studies

5.2 Analysis of repeated measures designs

5.2.1 Two-way ANOVA model

- Model
- Estimators
- Sum of squares and mean squares
- Statistical inference
 - Hypothesis testing
 - Confidence intervals

5.2.2 More complicated repeated measures design

- Two factors with repeated measures on one factor
- Two factors with repeated measures on both
- Split-plot design

For more discussion, see Chapter 8 here.

5.2.3 Longitudinal data analysis

We consider the rat growth data. Each rat is measured over 5 weeks. This type of data set is called longitudinal since the observations are taken over time. There is a covariate “mother’s weight” (X). The idea is to see how rat weights vary over time since birth. In another example, logarithm of CD4 counts are listed for patients on three different treatments over time. Goal is to investigate how CD4 counts change over time and if age has any effect on this change. Note that in the first example, the times at which measurement are taken are the same for all subjects. In the second case times may be different for different patients.

```
Rat.growth <- read.csv(file="./data/Growth.csv", header=TRUE, sep=",")
```

```
colorpicks = rainbow(n=length(unique(Rat.growth$rat)));
```

```
with(Rat.growth, plot(weight, week,type='p', pch=16, bty='l', main='Raw growth data', x
```

```
for(i in 1:length(unique(Rat.growth$rat))){
```

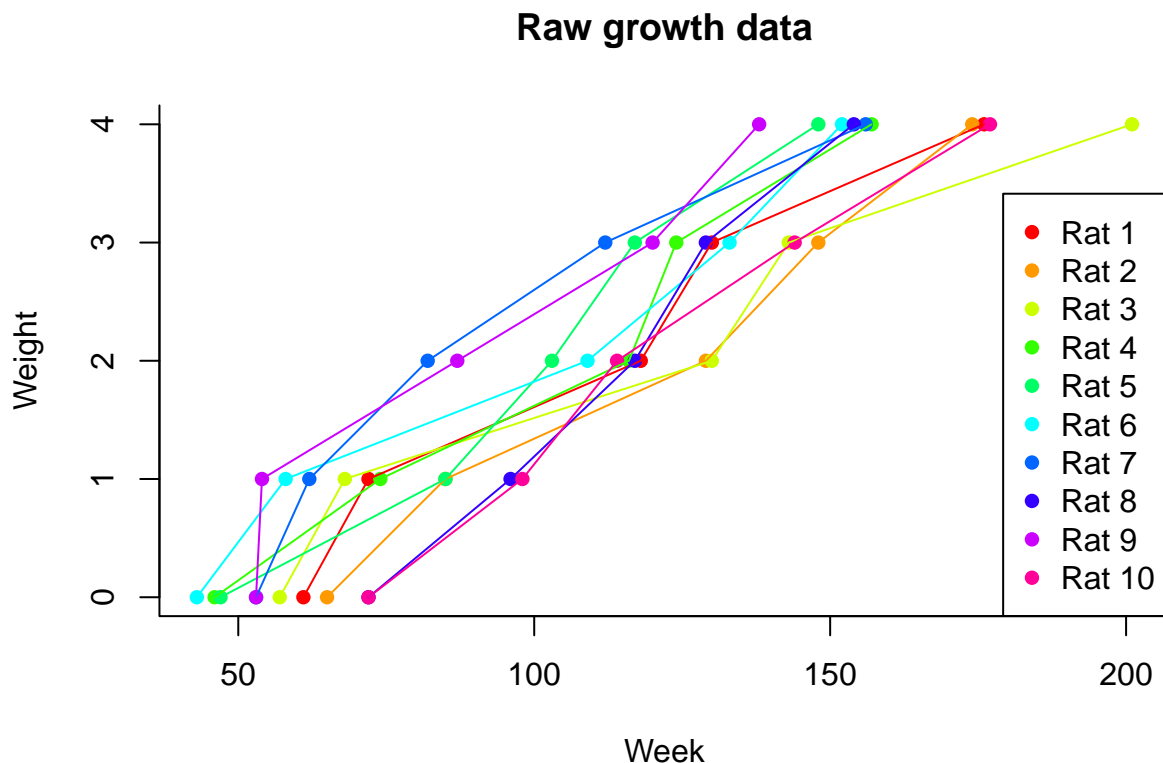
```
  one.rat=Rat.growth[Rat.growth$rat==i,]
```

```
  with(one.rat, lines(weight, week,col=colorpicks[rat])
```

```
}
```

```
# For more thoughts on visualization, see http://www.colbyimaging.com/wiki/statistics/
```

```
legend('bottomright', col=colorpicks, pch=c(16), legend=paste('Rat', unique(Rat.growth$
```



We consider several models to fit them in R. For more on the syntax of `lmer` see the vignette [here](#).

```
library(lme4)
lm1=lmer(weight~as.factor(week)+(1|rat),data=Rat.growth)
lm2=lmer(weight~week+X+(1|rat),data=Rat.growth)
lm3=lmer(weight~as.factor(week)+X+(1|rat),data=Rat.growth)
lm4=lmer(weight~week+X+(1|rat)+(0+week|rat),data=Rat.growth)
lm5=lmer(weight~week+I(week^2)+X+(1|rat),data=Rat.growth)
lm6=lmer(weight~week+I(week^2)+X+(1|rat)+(0+week|rat)+(0+week^2|rat),data=Rat.growth)

# For model selection, we can use AIC, BIC
AIC(logLik(lm1))

## [1] 380

# or
AIC(lm1,lm2)

##      df AIC
## lm1   7 380
## lm2   5 390

BIC(lm1,lm2,lm3,lm4,lm5,lm6)

##      df BIC
## lm1   7 393
## lm2   5 400
## lm3   8 389
## lm4   6 404
## lm5   6 401
## lm6   8 409
```


Chapter 6

Case-control Study

(In progress)

6.1 Case-control study

- Study design
- Sampling schemes
- Comparisons to other studies
 - randomized experiments
 - retrospective cohort studies
- Motivation for a case-control study
- Estimands in a case-control study

6.2 Logistic regression

- Logistic regression models
- Estimation via maximum likelihood
 - Simple case with analytic solution
 - Score function
 - Fisher information
- Statistical inference
 - Hypothesis testing
 - Confidence intervals
- Diagnostic plots
 - Residuals
 - Pearson residuals

6.3 Generalized linear model

- Basics of generalized linear models

- Practical use of GLM

Chapter 7

Observational Study

(Optional topic)

7.1 Causality in observational studies

- Features of observational studies
- Selection bias

7.2 Analysis with no latent confounding

- Assumptions
 - I.I.D.
 - Ignorability
 - Overlap
- Estimation
 - Stratification
 - Outcome regression
- Propensity score
 - Definition and key properties
 - Propensity scores: matching
 - Propensity scores: weighting
 - Doubly-robust regression
- Covariance balancing

7.3 Instrumental variable

- Definition and assumptions
- Key properties of IV
- Estimation

7.4 Missing data

- Missing mechanisms
- Multiple imputation

Chapter 8

Complex data

(Optional topic)

8.1 Data in the big data era

8.2 Useful methods

8.2.1 Penalized regression

- Motivation
- Loss function
- Ridge (ℓ_2 -penalty)
- Lasso (ℓ_1 -penalty)
- Other penalties

8.2.2 Model selection

- Motivation
- Stepwise selection
- Cross-validation

8.2.3 Other

- Dimension reduction
- False discovery rate
- Nonparametric models

Chapter 9

Project description

If a statistical method is employed, you need to clearly state the model and justify your choice.

9.1 Project 1: Project STAR I

9.1.1 Background

In this project, we study the dataset from a very influential randomized experiment. Tennessee Student/Teacher Achievement Ratio study (Project STAR) was conducted in the late 1980s to evaluate the effect of class size on test scores. This dataset has been used as a classic example in many textbooks and research papers. You are encouraged to read more about the experiment design and how others analyze this dataset. This document only provides a brief explanation of the dataset that suffices for this course project.

The study randomly assigned students to small classes, regular classes, and regular classes with a teacher's aide. In order to randomize properly, schools were enrolled only if they had enough studybody to have at least one class of each type. Once the schools were enrolled, students were randomly assigned to the three types of classes, and one teacher was randomly assigned to one class.

The dataset contains scaled scores for math and reading from kindergarten to 3rd grade. We will only examine the math scores in 1st grade in this project.

9.1.2 Tasks

1. Install the **AER** package and load the **STAR** dataset.
2. Explore this dataset and generate summary statistics (in forms of tables or plots) that you find informative, and explain them.

3. Write down a one-way ANOVA model to study the effects of class types on the math scaled scores. Explain your notation.
4. Explain why your model is appropriate for this task on this data set. You may want to include statistics and plots in your explanation.
5. Fit the model you choose in Task 3 and show your fits in the report.
6. Conduct model diagnostic and/or sensitivity analysis.
7. Test whether there is a difference in the math scaled score in 1st grade across students in different class types. Justify your choice of test.
8. Discuss whether you are able to make any causal statements based on your analysis.

9.2 Project 2: Project STAR II

9.3 Background

We will continue our study on Project STAR. In the previous project, we examine the effects on individual students. Here we consider each teacher as the individual unit. Moreover, we will also explore the longitudinal feature of this dataset and the fact that randomization happened within schools.

9.4 Tasks

1. Explore math scaled scores in the 1st with teachers as the unit. Generate summary statistics (in forms of tables or plots) that you find informative, and explain them.
2. Write down a two-way ANOVA model to study the effects of class types on math scaled scores in 1st grade, with the school indicator as the other factor. Explain your notation.
3. Explain why your model is appropriate for this task on this data set. You may want to include statistics and plots in your explanation.
4. Fit the model you choose in Task 2 and show your fits in the report.
5. Conduct model diagnostic and/or sensitivity analysis.
6. Test whether there is a difference in math scaled score in 1st grade across teachers in different class types. Justify your choice of test.
7. Discuss whether you are able to make any causal statements based on your analysis.
8. Is there any difference between the results from Project 2 compared to the results from Project 1?

In any of these tasks, if a statistical method is employed, you need to clearly state the model and justify your choice.

9.5 Project 3: US Traffic Fatalities

9.6 Background

The National Highway Traffic Safety Administration reported that there are 36,560 highway fatalities across US in 2018 (link). Alarmed by the high traffic fatalities, you wanted to use your knowledge and skills in statistics to explore measures that can potentially reduce the traffic fatalities.

You found out that traffic fatalities data for 48 US states from 1982 to 1988 are available, and well-cleaned, in an R package `AER`. You can see more description of the data set from the help file of the `fatalities` dataset (e.g., using `?Fatalities`).

You want to analyze this dataset to see if there are any of the variables that *caused* the reduction or increase of traffic fatalities. Note that the ultimate goal is to make suggestions to policy makers to take certain measures.

The beer tax is the tax on a case of beer, which is an available measure of state alcohol taxes more generally. The drinking age variable is a factor indicating whether the legal drinking age is 18, 19, or 20. The two binary punishment variables describe the state's minimum sentencing requirements for an initial drunk driving conviction.

Total vehicle miles traveled annually by state was obtained from the Department of Transportation. Personal income was obtained from the US Bureau of Economic Analysis, and the unemployment rate was obtained from the US Bureau of Labor Statistics

9.7 Tasks

1. Explore this dataset and generate summary statistics (in forms of tables or plots) that you find crucial for your own interest, or for convincing the policy makers.
2. Consider only the the full dataset from 1982 to 1988, propose a regression model to study whether having a mandatory jail sentence is associated with reduced traffic fatalities. In particular, you need to
 - a. write down your model,
 - b. state the assumptions required,
 - c. fit the model with appropriate methods,
 - d. conduct model diagnostics and/or sensitivity analysis.
 - e. and discuss causal interpretation of the proposed models.
3. Conclude your analysis results. You may want to test a hypothesis, construct a confidence interval, or draw a confidence band.
4. Explain the implications of your results to policy makers who know little about statistics. Make suggestions if you want to.

9.8 Hints

1. Take a look at the helpfile for this data set.
2. You may want to think about the causal interpretation (2.f) before finding an appropriate model (2.a).
3. Some key words here: longitudinal data, time-series data, propensity score, matching, instrument variable, and observational study. Not all key words lead to useful resources.
4. You may use additional data or studies to strengthen your argument in Task 4.

9.9 Project 4: Bank Marketing

9.10 Background

A Portuguese retail bank started a telemarketing campaign in 2008, aiming to subscribe new users to a long-term deposit. Information collected during the campaign was recorded in this data set. More information is available on the UCI machine learning repository ([link](#)) and the citations therein.

You are a consultant who is hired to study the retail banking market in Portuguese. Somehow you come to know such a dataset is publicly available. Naturally, you want to gain some insights of the market from this dataset, before conducting an expensive survey.

9.11 Tasks

1. Acquire the dataset from the UCI machine learning repository ([link](#)).
2. Pick an appropriate data set to study, and justify your decision.
3. Explore this dataset and generate summary statistics (in forms of tables or plots) that you find crucial for your clients to know.
4. Build a predictive model for whether a client will sign on to a long-term deposit. You will use logistic regression in this task. Specifically, you will
 - a. write down a proper logistic regression model,
 - b. fit the model,
 - c. evaluate the performance of the fitted model,
 - d. and conduct model diagnostic and/or sensitivity analysis.
5. Build another predictive model using a method of your choice, and compare its performance to the logistic regression.
6. Explain the gap in the performances, if any, to your supervisor, who knows statistics quite well and only believes in data and mathematics.

In any of these tasks, if a statistical method is employed, you need to clearly state the model and justify your choice.

9.12 Hints

1. Pay close attention to the dataset (Task 3) before building your prediction model.
2. You may use random forest in Task 5 if you have no other methods in mind.
3. You need to evaluate performance of each method in a statistically appropriate manner.
4. You may use methods built in other computing language and display the results in your report. Your code in other language should be made available for reproducibility.