

STA303/1004 - Intro to one-way ANOVA

January 9, 2018

Week 1 Topics

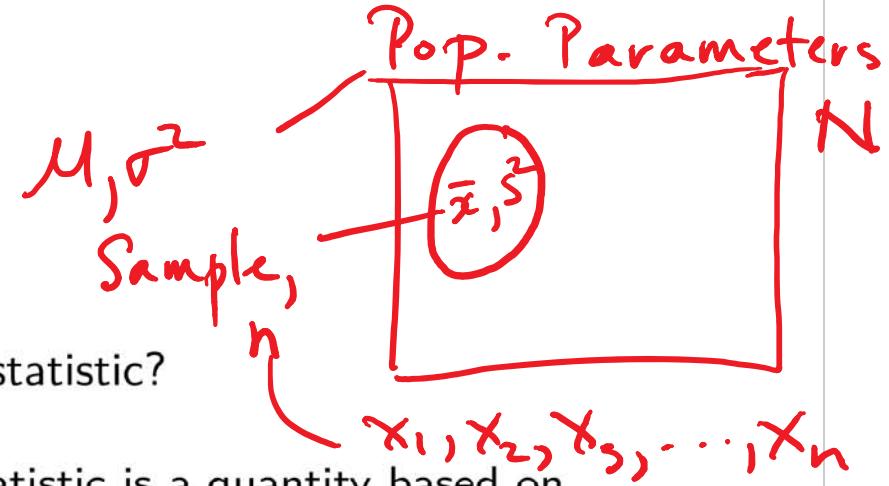
REVIEW

- Data summary: Five-number summary, Boxplots, *t*-tables
- Large-sample distribution theory: derived from Normal (T, Z, χ^2, F)
- Statistical inference: confidence interval, hypothesis tests, errors, power
- Normality Test, Equal variance test

T-TESTS

- One-sample t-test
- Paired t-test
- Two-sample t-test
- Non-parametric alternatives

Parameters and Statistics



What is the difference between a parameter and a statistic?

- ▶ A parameter is a population quantity and a statistic is a quantity based on a sample drawn from the population.

Example: The population of all adult (18+ years old) males in Toronto, Canada.

- ▶ Suppose that there are N adult males and the quantity of interest, y , is age.
- ▶ A sample of size n is drawn from this population.
- ▶ The population mean is $\mu = \sum_{i=1}^N y_i / N$.
- ▶ The sample mean is $\bar{y} = \sum_{i=1}^n y_i / n$.

The Normal Distribution

The density function of the normal distribution with mean μ and standard deviation σ is:

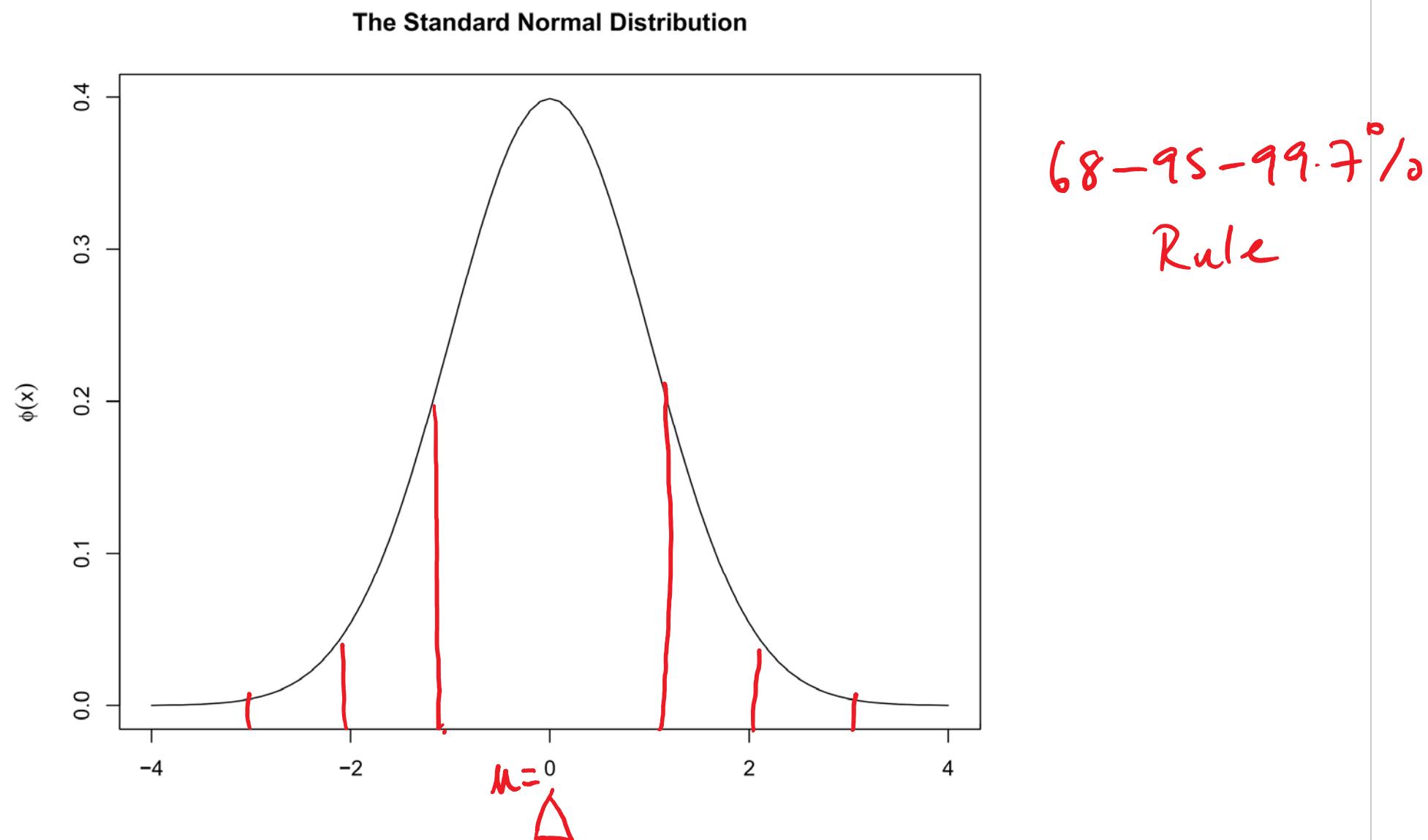
$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2} \left(\frac{x-\mu}{\sigma}\right)^2\right)$$

The cumulative distribution function (CDF) of a $N(0, 1)$ distribution,

$$\Phi(x) = P(X < x) = \int_{-\infty}^x \phi(x) dx$$

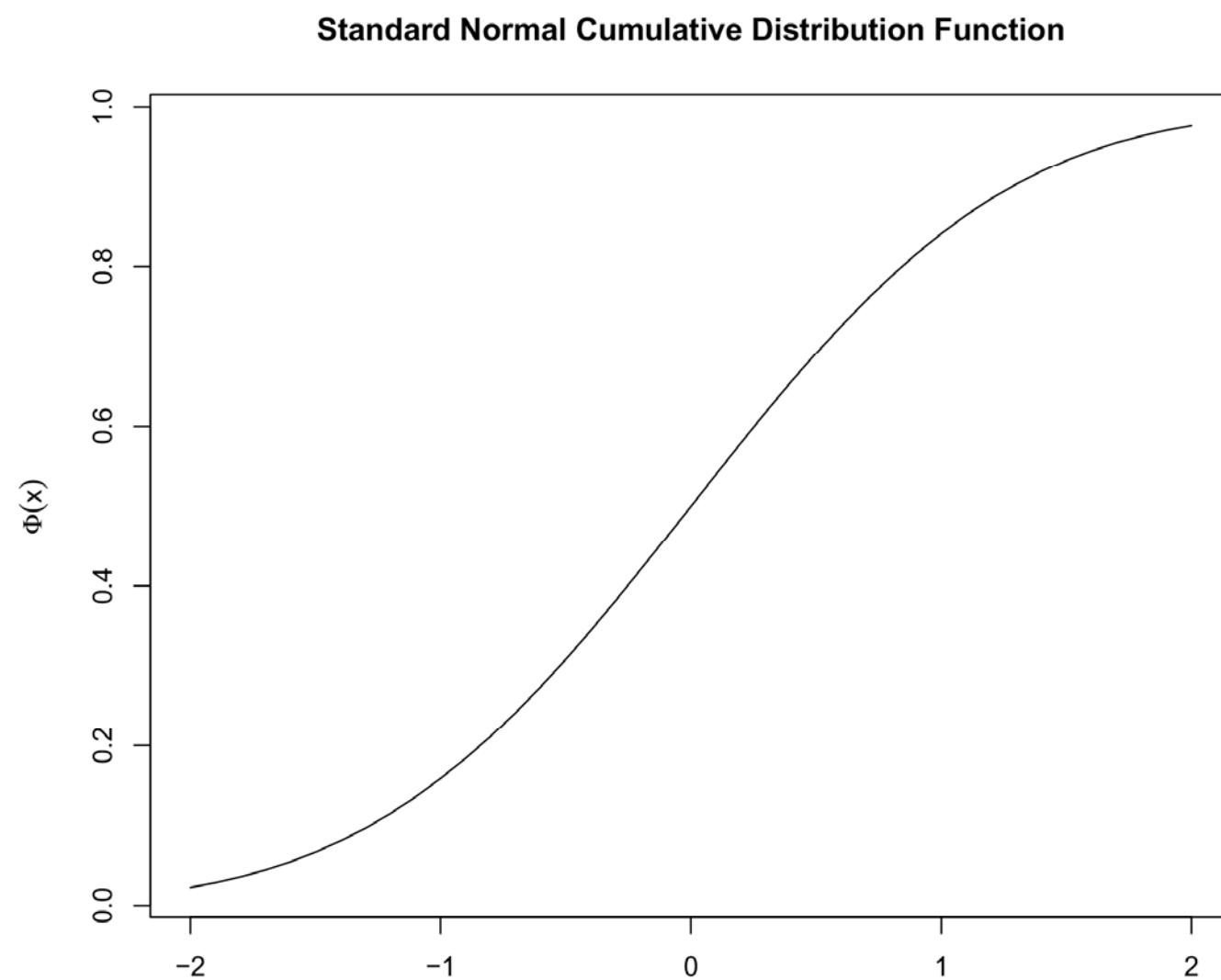
The Standard Normal Distribution

```
x <- seq(-4,4,by=0.1)
plot(x,dnorm(x),type="l",main = "The Standard Normal Distribution",
      ylab=expression(paste(phi(x))))
```



The Standard Normal CDF

```
plot(x <- seq(-2,2,by=0.1),pnorm(x),type="l",
      xlab="x",ylab=expression(paste(Phi(x))),
      main = "Standard Normal Cumulative Distribution Function")
```



The Normal and Standard Normal Distributions

A random variable X that follows a normal distribution with mean μ and variance σ^2 will be denoted by

$$X \sim N(\mu, \sigma^2).$$

If $X \sim N(\mu, \sigma^2)$ then

$$Z \sim N(0, 1),$$

where

$$Z = \frac{X - \mu}{\sigma}.$$

The Normal Distribution

$X \sim N(0, 1)$. Use R to find $P(-2 < X < 2)$.

```
pnorm(2,mean = 0,sd = sqrt(1))-pnorm(-2,mean = 0,sd = sqrt(1))  
## [1] 0.9544997
```

Normal Quantile-Quantile Plots

- used to visually assess Normality of a sample of measurements
- in R, use `qqnorm()` for the normal qq plot and `qqline()` to add the straight line.

Linear combination of IID Normal

If $X_i \sim N(\mu_i, \sigma_i^2)$ independently, then

$$V = a + \sum_{i=1}^n b_i X_i \sim N(a + \sum_{i=1}^n b_i \mu_i, \sum_{i=1}^n b_i^2 \sigma_i^2)$$

Chi-Square Distribution

Let X_1, X_2, \dots, X_n be independent and identically distributed random variables that have a $N(0, 1)$ distribution. The distribution of

$$\sum_{i=1}^n X_i^2, \sim \chi_n^2$$

$$z^2 \sim \chi_1^2$$

has a chi-square distribution on n degrees of freedom or χ_n^2 .

The mean of a χ_n^2 is n with variance $2n$.

$$H_0: \sigma^2 = \sigma_0^2 \quad (\sigma = \sigma_0)$$

Chi-Square Distribution

Let X_1, X_2, \dots, X_n be independent with a $N(\mu, \sigma^2)$ distribution. What is the distribution of the sample variance $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$?

$$\frac{(n-1) S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Sample variance

t Distribution

$$X \perp W$$

If $X \sim N(0, 1)$ and $W \sim \chi_n^2$ then the distribution of $\frac{X}{\sqrt{W/n}}$ has a t distribution on n degrees of freedom or $\frac{X}{\sqrt{W/n}} \sim t_n$.

$$\frac{X}{\sqrt{W/n}}$$

$$T_{df} \xrightarrow{\mathcal{D}} z$$

t Distribution

Let X_1, X_2, \dots is an independent sequence of identically distributed random variables that have a $N(0, 1)$ distribution. What is the distribution of

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n-1}}}$$

where $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$?

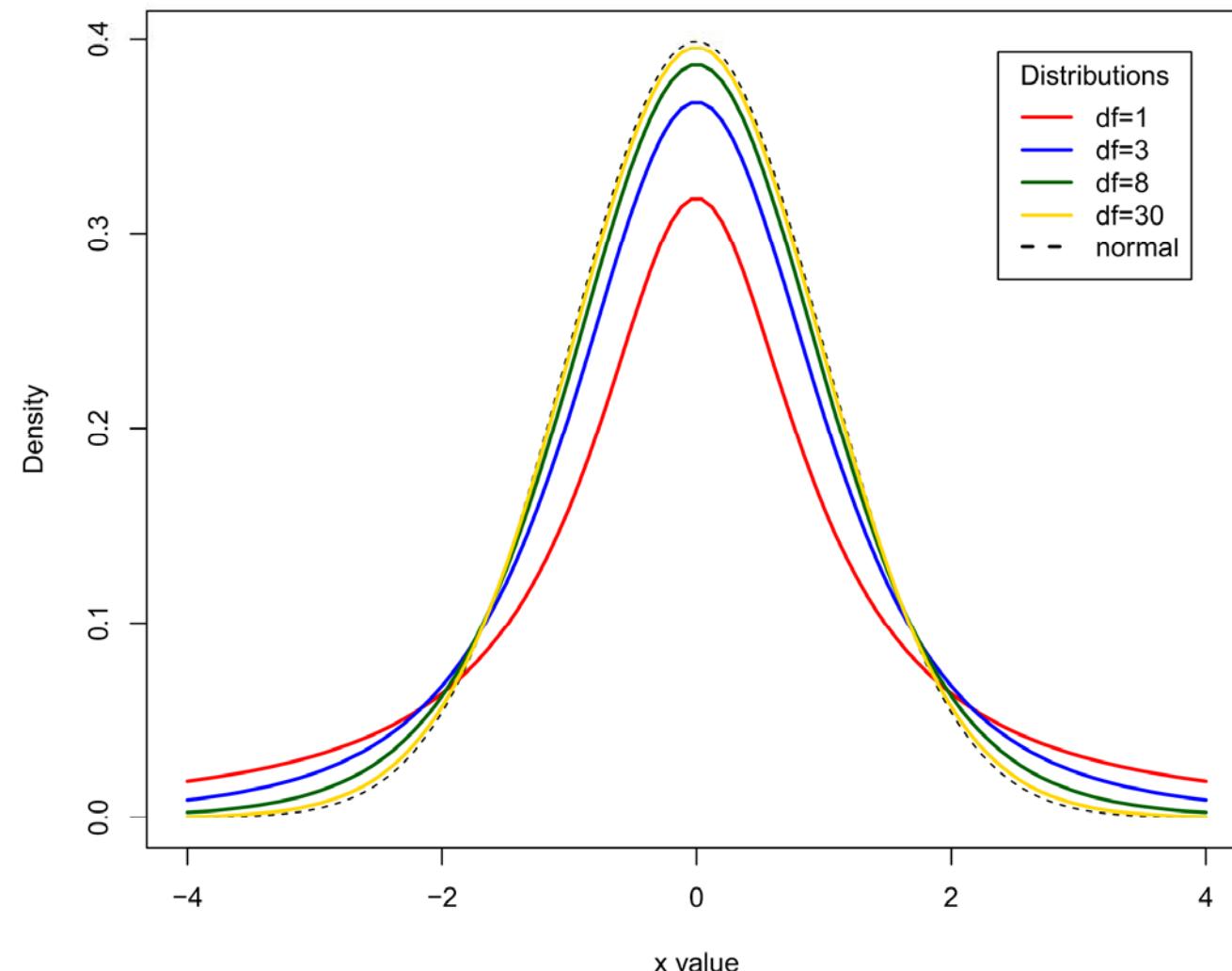
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{(n-1)S^2}{\sigma^2/(n-1)}}}$$

$$\sqrt{\chi^2_{n-1} / (n-1)}$$

$$\sim t_{n-1}$$

t Distribution

Comparison of t Distributions



F Distribution

Let $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ be independent. The distribution of

$$W = \frac{X/m}{Y/n} \sim F_{m,n},$$

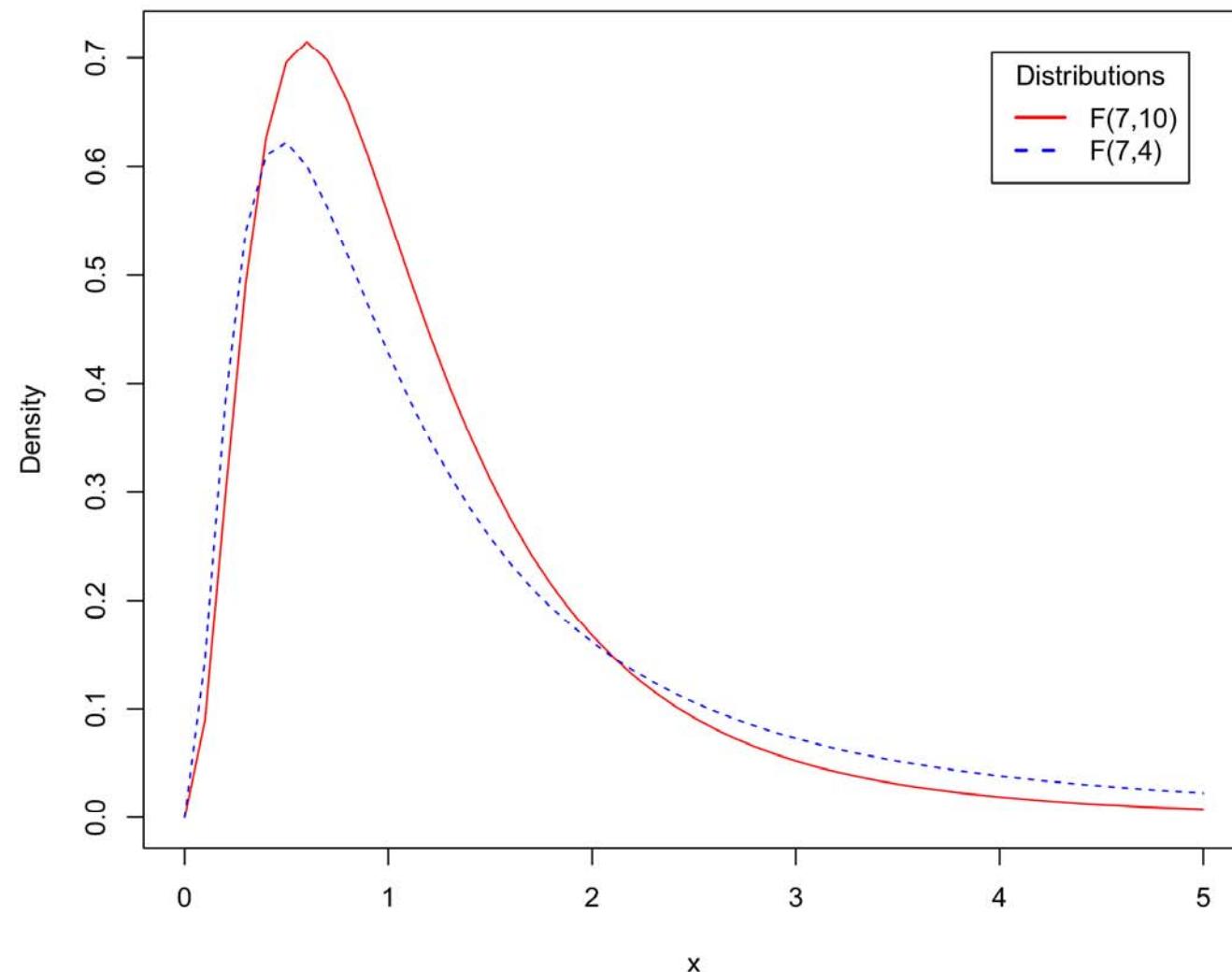
where $F_{m,n}$ denotes the F distribution on m, n degrees of freedom. The F distribution is right skewed (see graph below). For $n > 2$, $E(W) = n/(n - 2)$. It also follows that the square of a t_n random variable follows an $F_{1,n}$.

$$\frac{T^2}{df} \stackrel{D}{=} F_{1,df}$$

$$\begin{aligned} \chi^2: H_0: \sigma^2 &= \sigma_0^2 \\ F: H_0: \frac{\sigma_1^2}{\sigma_2^2} &= 1 \end{aligned}$$

F Distribution

F Distributions



The Sample Mean

If $X_1, \dots, X_n \sim_{iid} N(\mu, \sigma^2)$ then

- ▶ $\bar{X} \sim N(\mu, \sigma^2/n)$
- ▶ $S^2 = \sum(X - \bar{X})^2/(n - 1)$ and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

- ▶ $\bar{X} \perp S^2$ and

- ▶

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \boxed{\frac{\bar{X} - \mu}{S/\sqrt{n}}} \sim t_{n-1}$$

Simple Linear Regression

A simple linear regression model is obtained by estimating the intercept and slope in the equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \dots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$. The values of β_0, β_1 that minimize the sum of squares

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2,$$

are called the least squares estimators. They are given by:

- ▶ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- ▶ $\hat{\beta}_1 = r \frac{S_y}{S_x}$

r is the correlation between y and x , and S_x, S_y are the sample standard deviations of x and y respectively.

Case Study 1: The Spock Conspiracy Trial

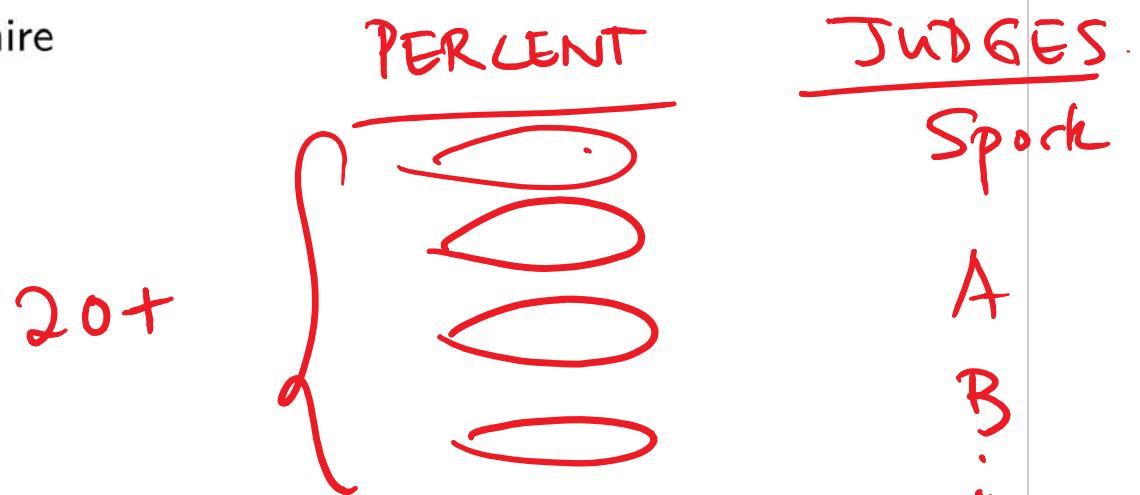
- ▶ Boston, 1968
- ▶ Dr. Benjamin Spock (paediatrician and author) on trial for conspiring to violate the Selective Service Act.
- ▶ Accused of encouraging people to dodge military draft by his books that advised on how mothers should raise children.
- ▶ Spock's jury had NO women.

Q: Is there evidence of gender bias in the jury selection for Spock's trial?

Case Study 1: Jury selection

- ▶ 300 names selected at random from city directory
- ▶ 35 to 200 jurors randomly selected (this group is called the venire)
- ▶ Then non-random selection or exclusion of jurors from the venire by both defence and prosecution
- ▶ For Spock's trial, only 1 woman in the venire but she was then dismissed by prosecution
- ▶ Defence argued that Spock's judge had history of women being underrepresented on his venires.
- ▶ Compared composition of recent venires of 6 other judges with that of Spock's judge
- ▶ Data: percent of women in each venire

venire
()

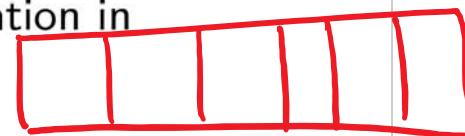


Case Study 1: Two Key Questions

- ▶ Q1. Is there evidence that women are underrepresented on Spock's judge's venires when compared to other judges?
- ▶ Q2. Is there evidence that there are differences in women's representation in venires of the other 6 judges?
- ▶ Q: Conduct the relevant hypothesis test to answer Q1. Include the necessary assumptions, justifications and elements of a hypothesis test.
What is your conclusion in plain English?

Spock's vs Others.

Others.



Case Study 1: The Spock Conspiracy Trial Data

The data is shown below.

```
#Juries data  
juries<-read.csv(  
  "/Users/Shivon/STA303_1002/LectureNotes/Lec1/juries.csv", header=T)
```

```
attach(juries)
```

```
#head(juries)
```

```
PERCENT
```

dim
length

```
## [1] 6.4 8.7 13.3 13.6 15.0 15.2 17.7 18.6 23.1 16.8 30.8 33.6 40.  
## [15] 27.0 28.9 32.0 32.7 35.5 45.6 21.0 23.4 27.5 27.5 30.5 31.9 32.  
## [29] 33.8 24.3 29.7 17.7 19.7 21.5 27.9 34.8 40.2 16.5 20.7 23.5 26.  
## [43] 29.5 29.8 31.9 36.2
```

JUDGE

```
## [1] SPOCKS  
## [11] A A A A B B B B B  
## [21] C C C C C C C C C  
## [31] D E E E E E E F F  
## [41] F F F F F F  
## Levels: A B C D E F SPOCKS
```

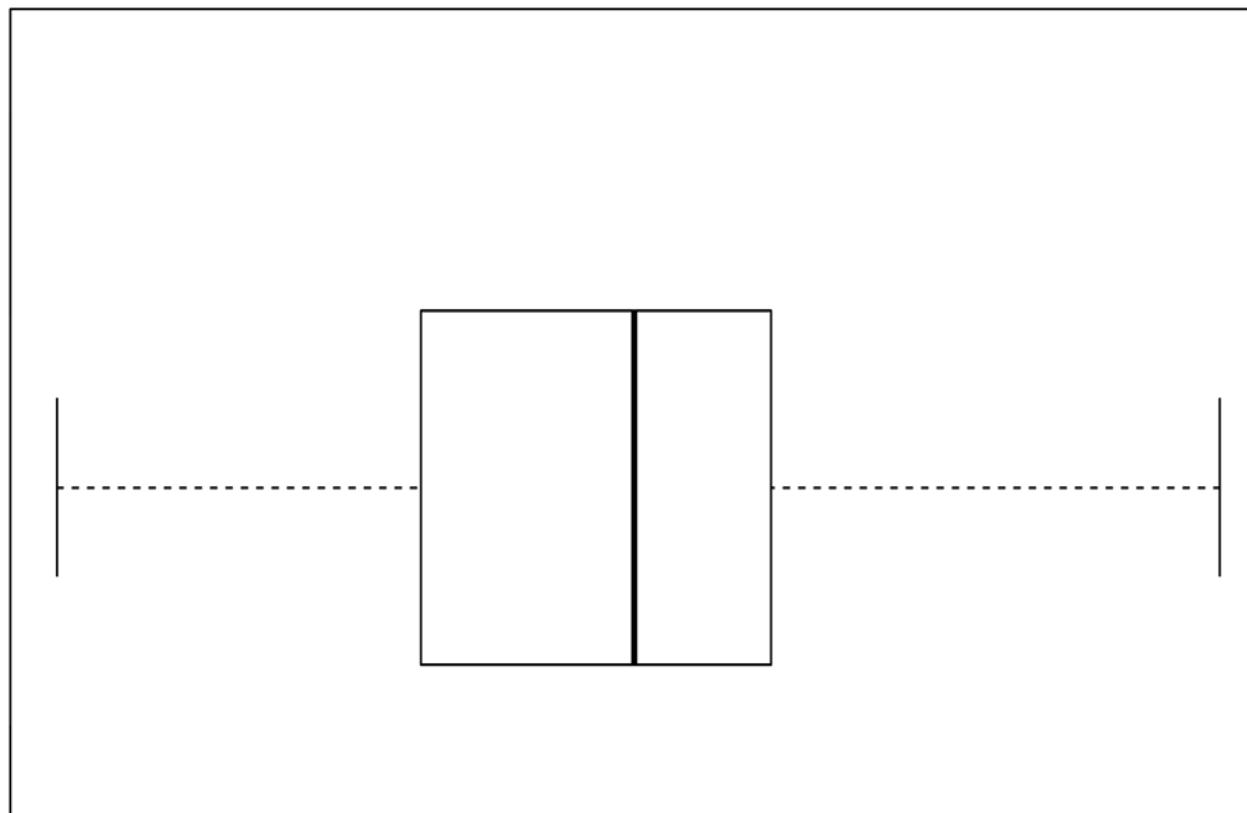
Case Study 1: Data summary

```
summary(PERCENT)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.    Max.  
##     6.40   19.95  27.50  26.58  32.38  48.90
```

```
boxplot(PERCENT, horizontal=T, main="Percent of women")
```

Percent of women *on all venues*



S-num summary

$$IQR = Q_3 - Q_1$$

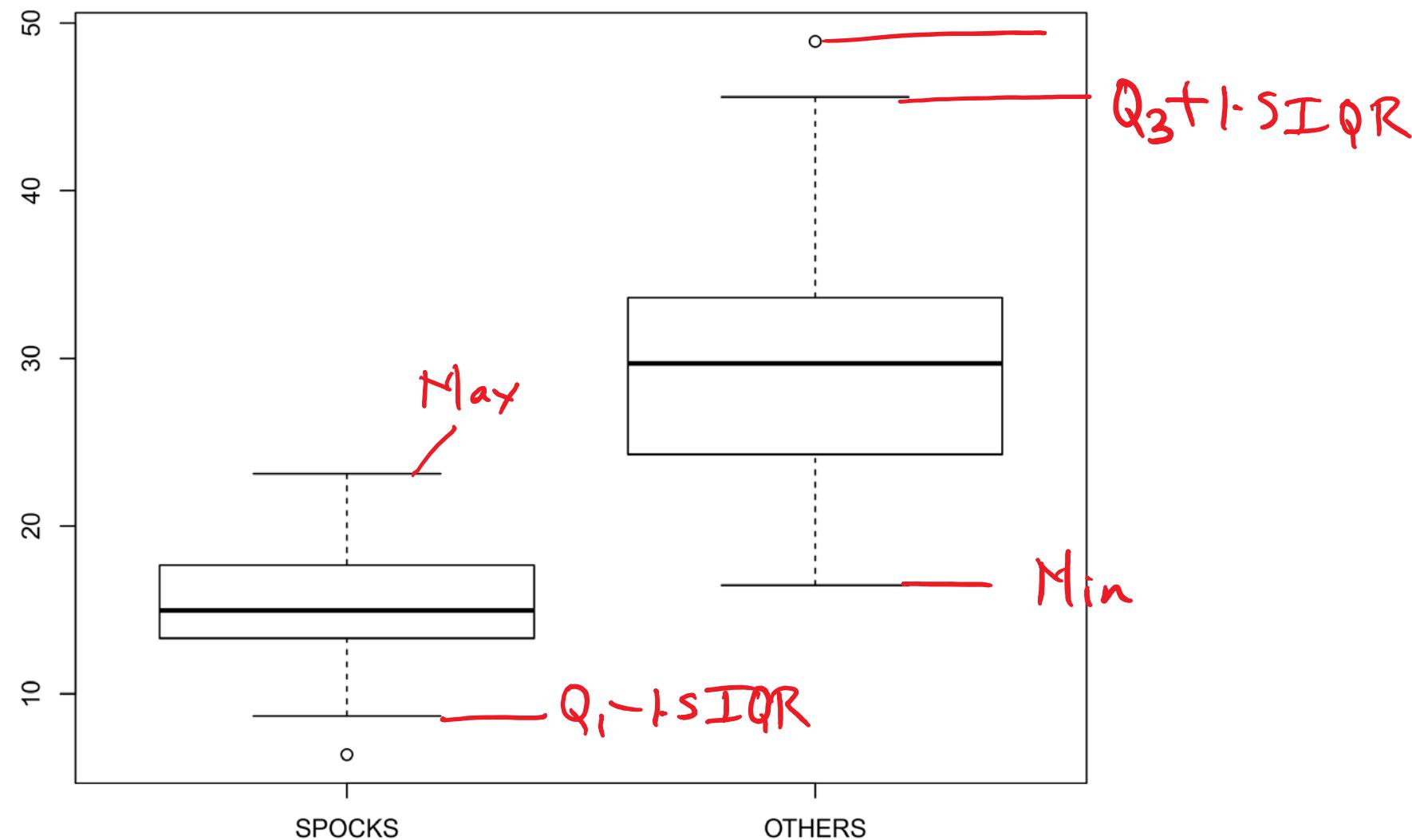
(robust measure
of spread).

middle 50%

Outlier Rule:
 $1.5 \times IQR$

Case Study 1: Two Sample t-tests

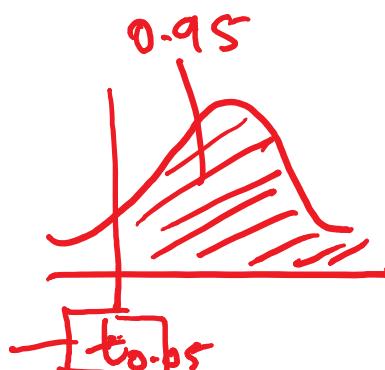
```
groupS<-PERCENT [JUDGE=="SPOCKS"]  
groupNS<-PERCENT [JUDGE!="SPOCKS"]  
boxplot(groupS, groupNS,xlab="JUDGE",names=c("SPOCKS", "OTHERS"))
```



Case Study 1: One Sample t-test

```
#one sample t test
t.test(PERCENT, mu=50)
```

```
##
## One Sample t-test
##
## data: PERCENT
## t = -17.303, df = 45, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
## 23.85675 29.30847
## sample estimates:
## mean of x
## 26.58261
```



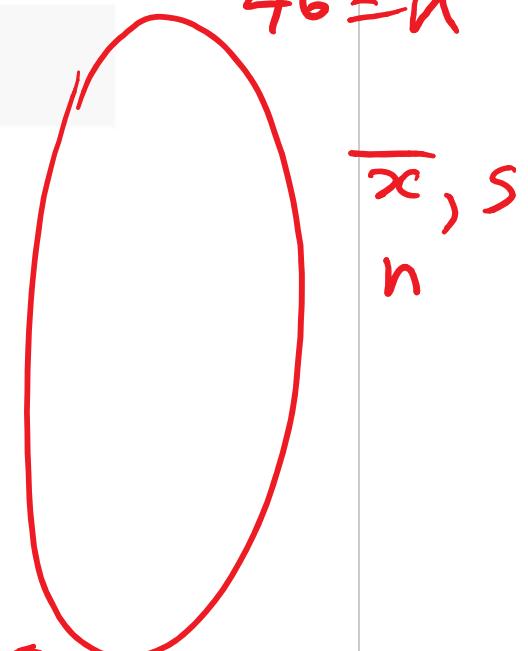
(Test Assumptions) ←
 1. Random obs.
 2. Approx. Normal pop.

μ : true % of women on ventres

$$46 = n$$

$$H_0: \mu = 50\%$$

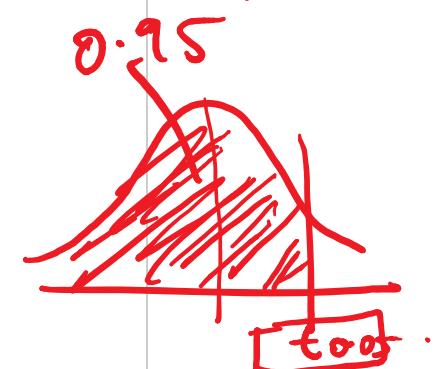
$$H_a: \mu \neq 50\%$$



$$t = \frac{\bar{x} - 50}{s/\sqrt{n}} \sim t_{45}$$

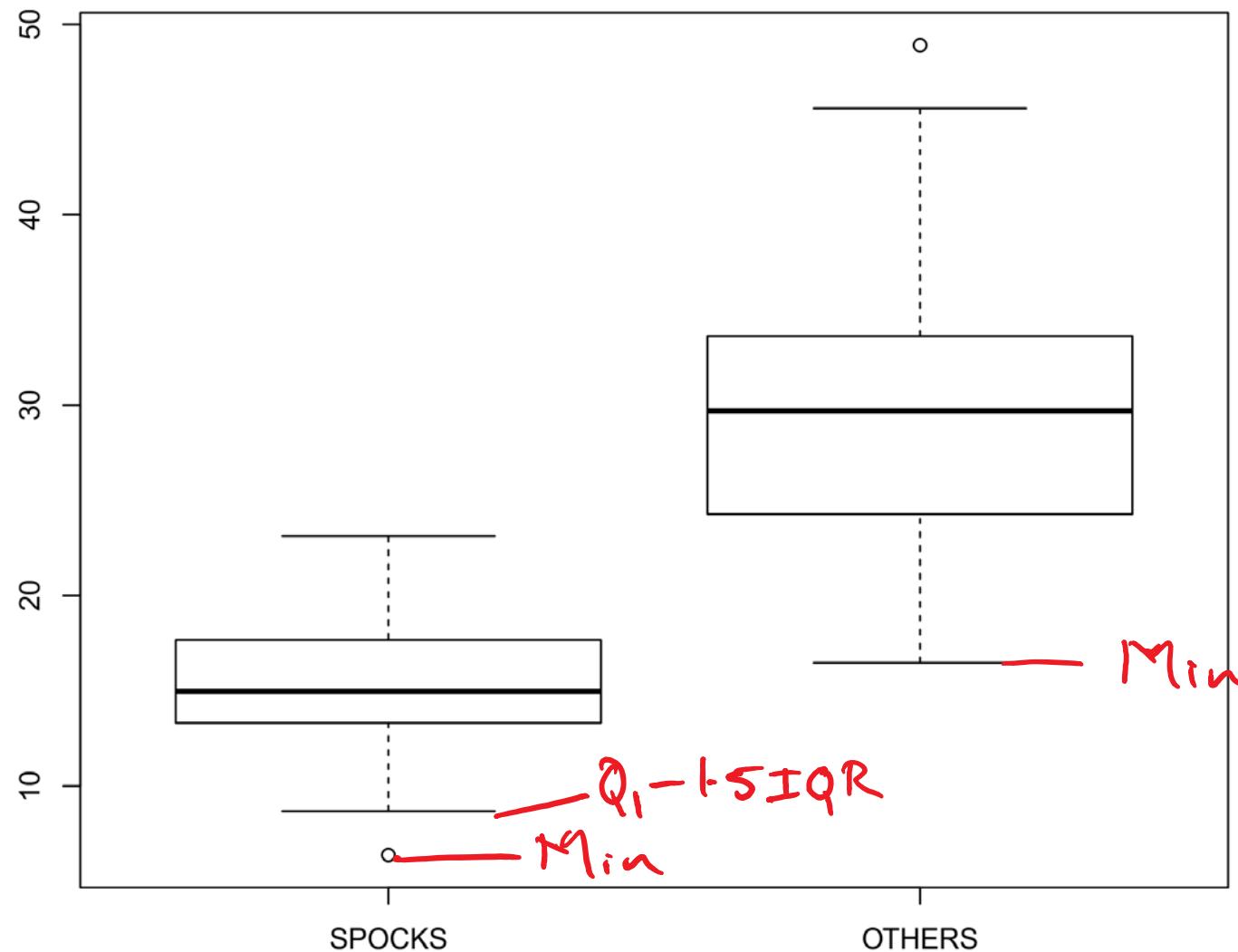
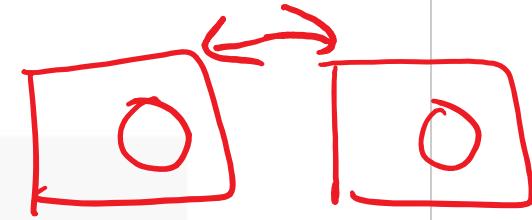
$$P\text{-value} = 2 P(T_{45} > |-17.303|)$$

$$C.I.: \bar{x} \pm t_{45, 0.025} \frac{s}{\sqrt{n}}$$



Case Study 1: Two Sample t-tests

```
groupS<-PERCENT [JUDGE=="SPOCKS"]  
groupNS<-PERCENT [JUDGE!="SPOCKS"]  
boxplot(groupS, groupNS,xlab="JUDGE",names=c("SPOCKS", "OTHERS"))
```



1. Independent samples

2. Samples are from approx. Normal populations.

3. $\sigma_1 = \sigma_2$
of group 1
of group 2

Case Study 1: Checking equal variance assumption

```
var(groupS)
```

```
## [1] 25.38945
```

```
var(groupNS)
```

```
## [1] 55.21632
```

#Rule of Thumb

max(var(groupS), var(groupNS)) / min(var(groupS), var(groupNS)) > 4 ?

```
## [1] 2.174775
```

max(sd(groupS), sd(groupNS)) / min(sd(groupS), sd(groupNS)) > 2 ?

```
## [1] 1.474712
```

Case Study 1: Checking equal variance assumption

```
#F Test of Equal variances  
var.test(groupS, groupNS)
```

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \text{or} \quad \sigma_1^2 = \sigma_2^2$$

```
##  
## F test to compare two variances  
##  
## data: groupS and groupNS  
## F = 0.45982, num df = 8, denom df = 36, p-value = 0.2482  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.1789822 1.7739665  
## sample estimates:  
## ratio of variances  
## 0.4598178
```

- Assume equal variances

Case Study 1: Two sample (unpooled) t-tests

```
#Welch-Satterthwaite (Unpooled)  
t.test(groupS, groupNS, var.equal=F)
```

Assume $\sigma_1 \neq \sigma_2$

```
##  
## Welch Two Sample t-test  
##  
## data: groupS and groupNS  
## t = -7.1597, df = 17.608, p-value = 1.303e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -19.23999 -10.49935  
## sample estimates:  
## mean of x mean of y  
## 14.62222 29.49189
```

$H_0: \mu_S = \mu_O \text{ or } \mu_S - \mu_O = 0$

Case Study 1: Pooled t-test

```
#Pooled
t.test(groupS, groupNS, var.equal=T)
```

```
##
## Two Sample t-test
##
## data: groupS and groupNS
## t = -5.6697, df = 44, p-value = 1.03e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -20.155294 -9.584045
## sample estimates:
## mean of x mean of y
## 14.62222 29.49189
```

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Assume $\sigma_1 = \sigma_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Conc: There is evidence at 5% level that the % women on Spock's judges' venires is less than that of other judges.
($p=0.00000103$)

Case Study 1: Paired t-test

```
#Paired  
t.test(groupS, groupNS, paired=TRUE)  
  
## Error in complete.cases(x, y): not all arguments have the same length
```

Case Study 1: Pooled t-test (Left tailed)

```
#Left-tailed Pooled  
t.test(groupS,groupNS,alternative="less",var.equal=TRUE)
```

```
##  
## Two Sample t-test  
##  
## data: groupS and groupNS  
## t = -5.6697, df = 44, p-value = 5.148e-07  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
##       -Inf -10.463  
## sample estimates:  
## mean of x mean of y  
## 14.62222 29.49189
```

$$H_0: \mu_s \leq \mu_d$$

Case Study 1: Simple Linear Regression Approach

```
X=c(rep(1,length(groupS)), rep(0,length(groupNS)))
Y=PERCENT; model1<-lm(Y~X); summary(model1)

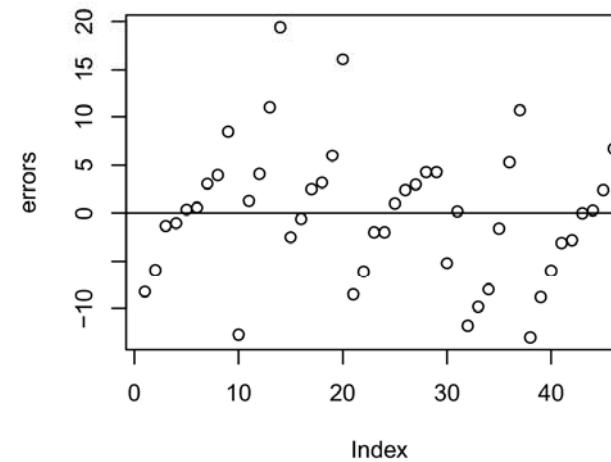
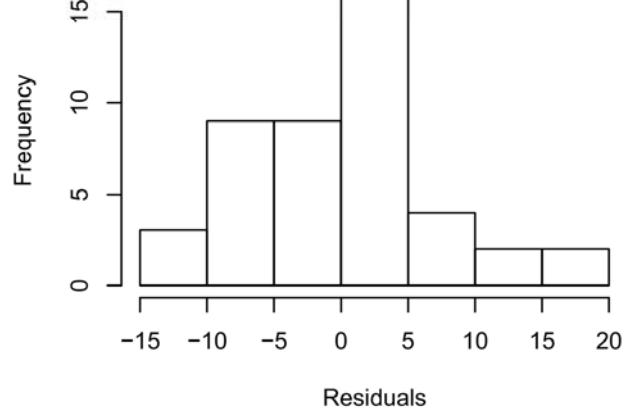
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9919  -4.6669   0.2581   3.7854  19.4081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.492     1.160   25.42 < 2e-16 ***
## X          -14.870     2.623   -5.67 1.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.056 on 44 degrees of freedom
## Multiple R-squared:  0.4222, Adjusted R-squared:  0.409
## F-statistic: 32.15 on 1 and 44 DF,  p-value: 1.03e-06
```

Case Study 1: Regression diagnostics

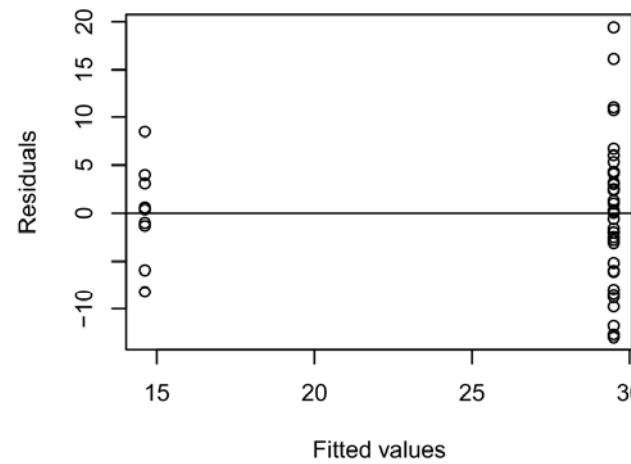
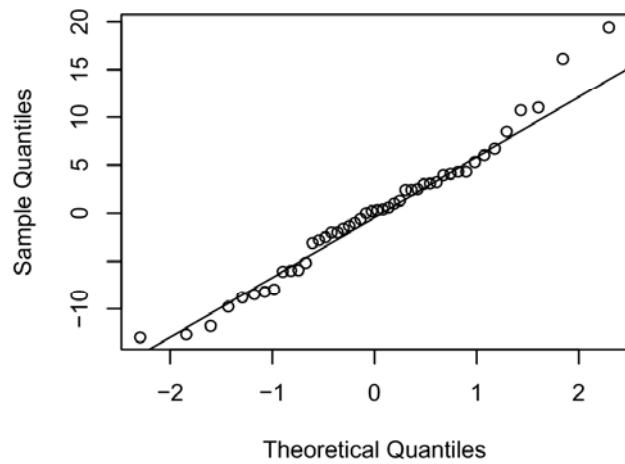
```
yhats=fitted(model1)
errors=residuals(model1)
# par(mfrow=c(2,2)) #partition plot window
# #plot (1,1)- histogram of residuals
# hist(errors, xlab="Residuals", breaks=5)
# #plot(1,2)- residuals vs index(time) with zero line
# # plot(errors)
# abline(0,0)
# #plot(2,1)-normal qq plot of residuals with qqline
# qqnorm(errors)
# qqline(errors)
# #plot(2,2)-residuals vs fitted values with zero line
# plot(yhats, errors, xlab="Fitted values", ylab="Residuals")
# abline(0,0)
```

Case Study 1: Regression diagnostics

Histogram of errors



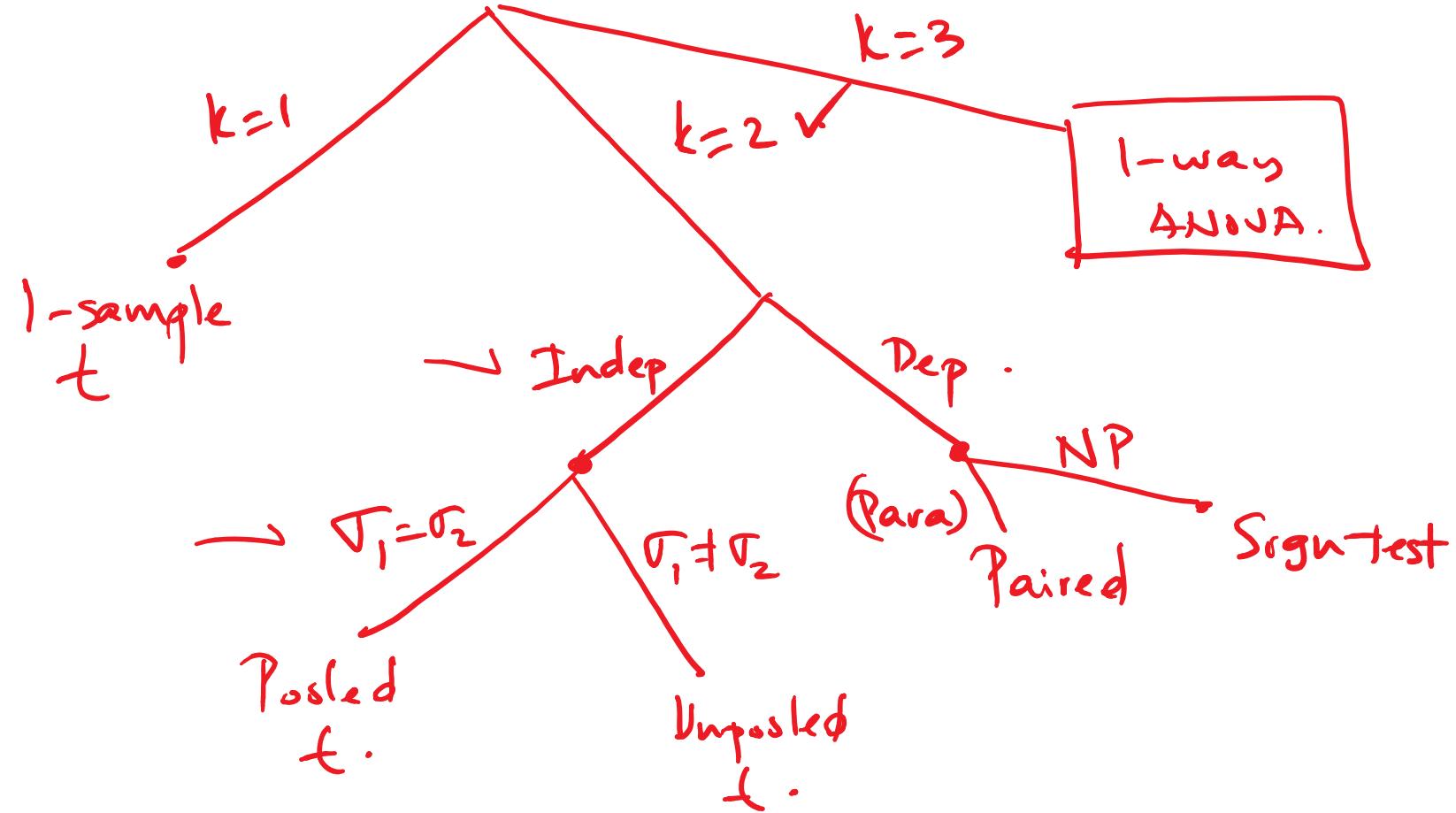
Normal Q-Q Plot



Case Study 1: One-way ANOVA approach

```
#ANOVA approach  
anova(model1)
```

```
## Analysis of Variance Table  
##  
## Response: Y  
##           Df Sum Sq Mean Sq F value    Pr(>F)  
## X          1 1600.6 1600.62  32.145 1.03e-06 ***  
## Residuals 44 2190.9   49.79  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



January 11, 2018

Intro to 1-way ANOVA

Week 1 Topics

REVIEW

- Data summary: Five-number summary, Boxplots
- Large-sample distribution theory: derived from Normal
- Statistical inference: confidence interval, hypothesis tests, errors, power
 - Normality Test, Equal variance test

T-TESTS

- One-sample t-test
- Paired t-test
- Two-sample t-test
- Non-parametric alternatives

Parameters and Statistics

What is the difference between a parameter and a statistic?

- ▶ A parameter is a population quantity and a statistic is a quantity based on a sample drawn from the population.

Example: The population of all adult (18+ years old) males in Toronto, Canada.

- ▶ Suppose that there are N adult males and the quantity of interest, y , is age.
- ▶ A sample of size n is drawn from this population.
- ▶ The population mean is $\mu = \sum_{i=1}^N y_i / N$.
- ▶ The sample mean is $\bar{y} = \sum_{i=1}^n y_i / n$.

The Normal Distribution

The density function of the normal distribution with mean μ and standard deviation σ is:

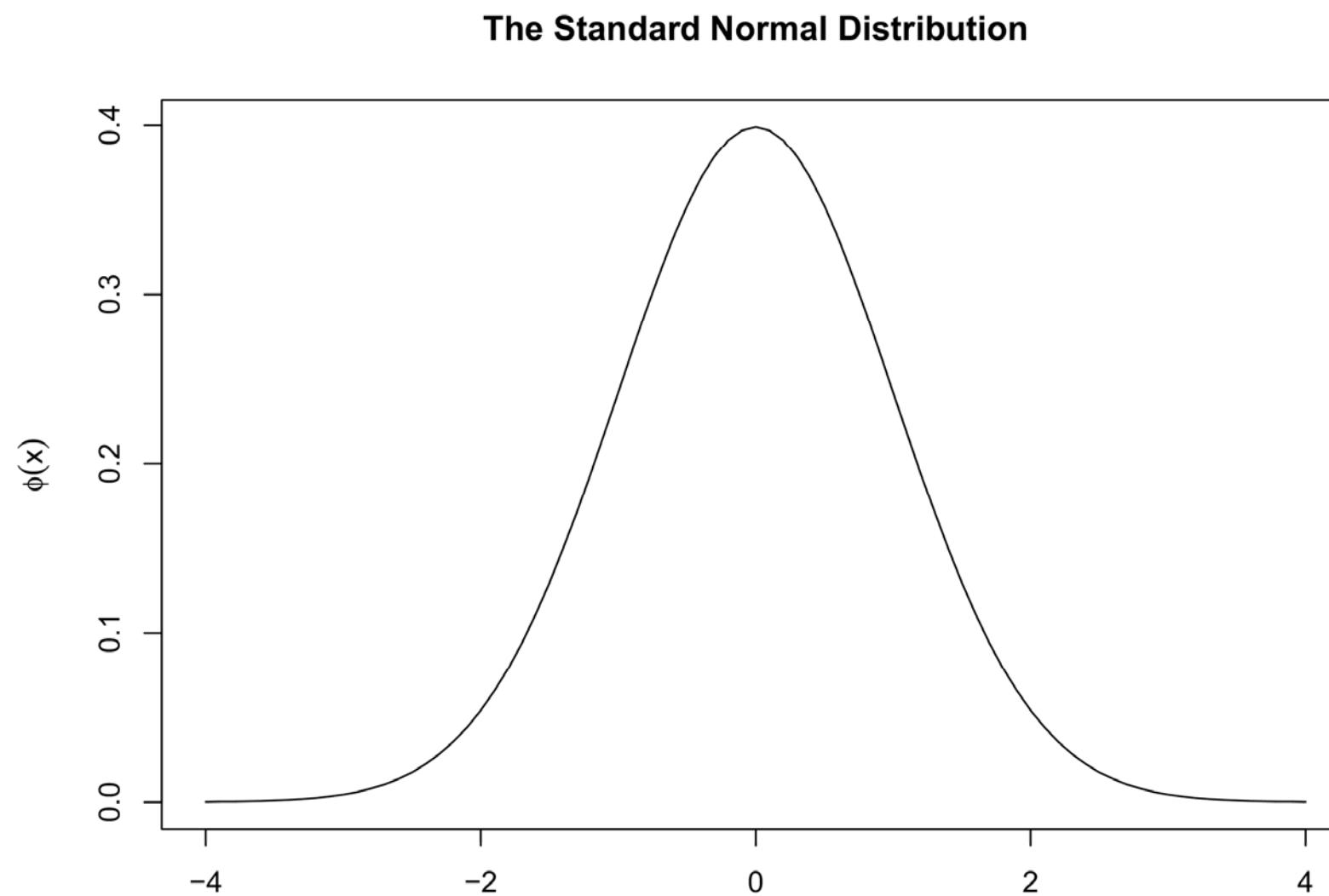
$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2} \left(\frac{x-\mu}{\sigma}\right)^2\right)$$

The cumulative distribution function (CDF) of a $N(0, 1)$ distribution,

$$\Phi(x) = P(X < x) = \int_{-\infty}^x \phi(x) dx$$

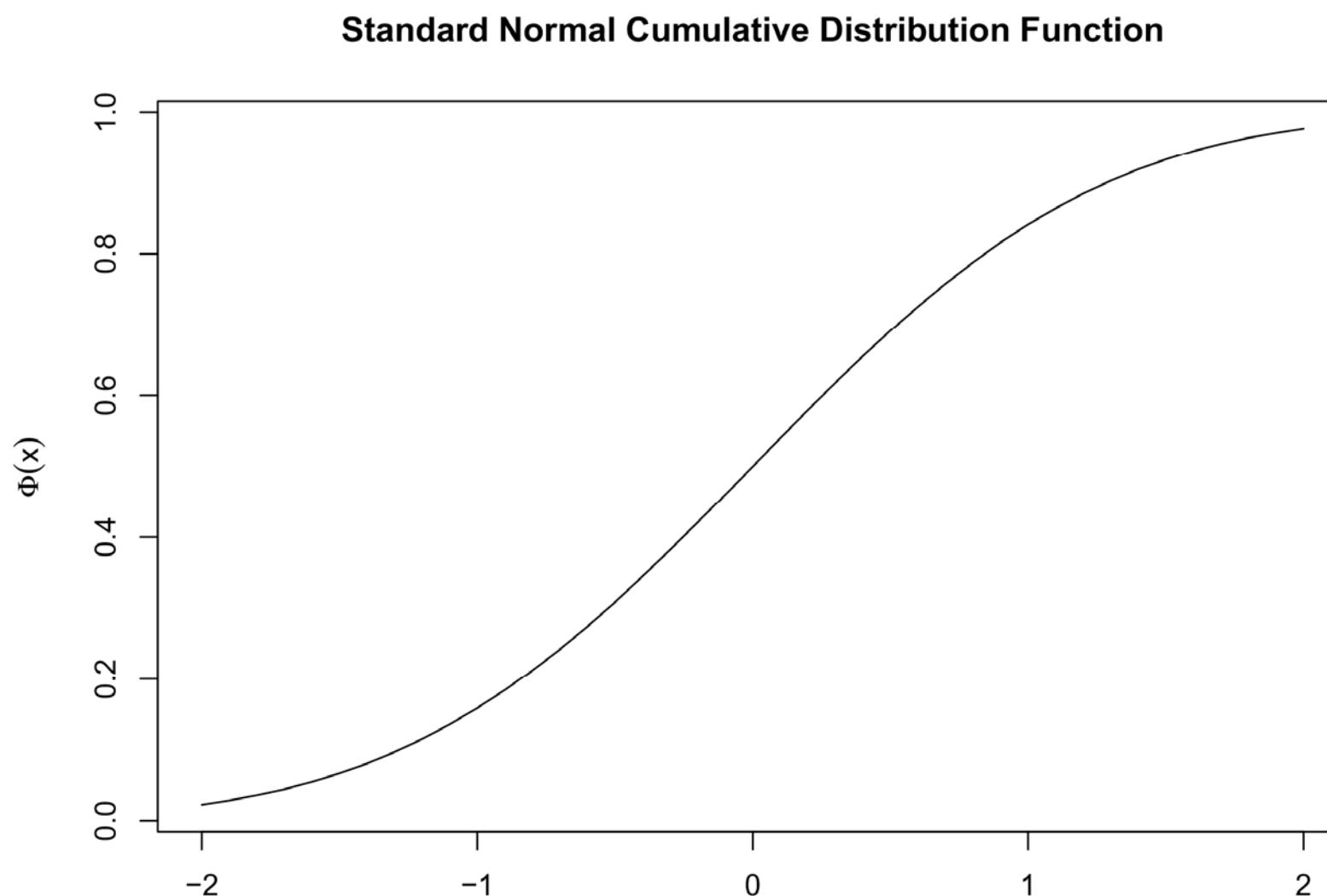
The Standard Normal Distribution

```
x <- seq(-4,4,by=0.1)
plot(x,dnorm(x),type="l",main = "The Standard Normal Distribution",
      ylab=expression(paste(phi(x))))
```



The Standard Normal CDF

```
plot(x <- seq(-2,2,by=0.1),pnorm(x),type="l",
      xlab="x",ylab=expression(paste(Phi(x))),
      main = "Standard Normal Cumulative Distribution Function")
```



The Normal and Standard Normal Distributions

A random variable X that follows a normal distribution with mean μ and variance σ^2 will be denoted by

$$X \sim N(\mu, \sigma^2).$$

If $X \sim N(\mu, \sigma^2)$ then

$$Z \sim N(0, 1),$$

where

$$Z = \frac{X - \mu}{\sigma}.$$

The Normal Distribution

$X \sim N(0, 1)$. Use R to find $P(-2 < X < 2)$.

```
pnorm(2,mean = 0,sd = sqrt(1))-pnorm(-2,mean = 0,sd = sqrt(1))  
## [1] 0.9544997
```

Normal Quantile-Quantile Plots

- used to visually assess Normality of a sample of measurements
- in R, use `qqnorm()` for the normal qq plot and `qqline()` to add the straight line.

Linear combination of independent Normals

If $X_i \sim N(\mu_i, \sigma_i^2)$ independently, then

$$V = a + \sum_1^n b_i X_i \sim N\left(a + \sum_1^n b_i \mu_i, \sum_1^n b_i^2 \sigma_i^2\right)$$

Chi-Square Distribution

Let X_1, X_2, \dots, X_n be independent and identically distributed random variables that have a $N(0, 1)$ distribution. The distribution of

$$\sum_{i=1}^n X_i^2,$$

has a chi-square distribution on n degrees of freedom or χ_n^2 .

The mean of a χ_n^2 is n with variance $2n$.

Chi-Square Distribution

Let X_1, X_2, \dots, X_n be independent with a $N(\mu, \sigma^2)$ distribution. What is the distribution of the sample variance $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$?

t Distribution

If $X \sim N(0, 1)$ and $W \sim \chi_n^2$ then the distribution of $\frac{X}{\sqrt{W/n}}$ has a t distribution on n degrees of freedom or $\frac{X}{\sqrt{W/n}} \sim t_n$.

t Distribution

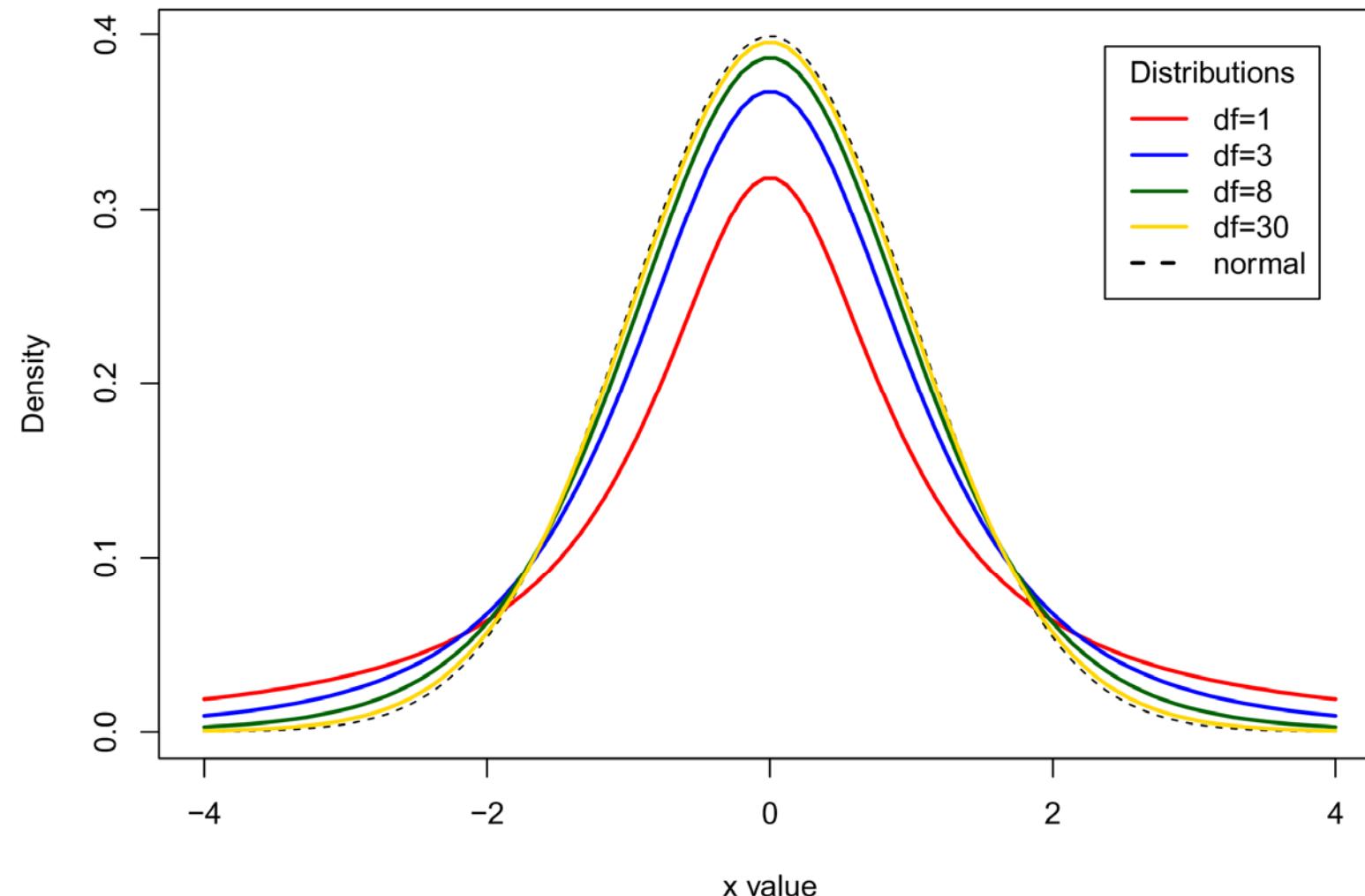
Let X_1, X_2, \dots is an independent sequence of identically distributed random variables that have a $N(0, 1)$ distribution. What is the distribution of

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

where $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$?

t Distribution

Comparison of t Distributions



F Distribution

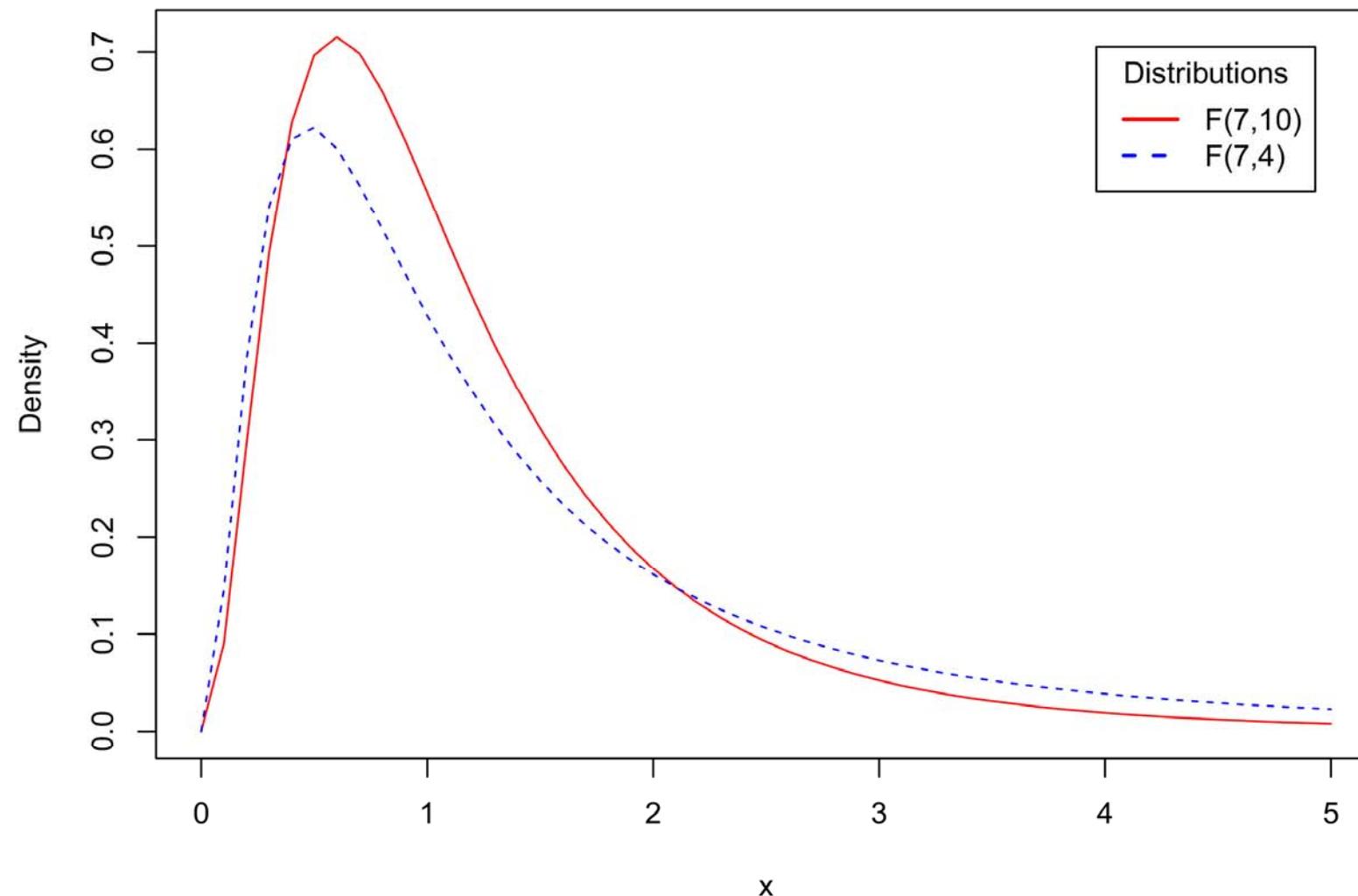
Let $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ be independent. The distribution of

$$W = \frac{X/m}{Y/n} \sim F_{m,n},$$

where $F_{m,n}$ denotes the F distribution on m, n degrees of freedom. The F distribution is right skewed (see graph below). For $n > 2$, $E(W) = n/(n - 2)$. It also follows that the square of a t_n random variable follows an $F_{1,n}$.

F Distribution

F Distributions



The Sample Mean

If $X_1, \dots, X_n \sim_{iid} N(\mu, \sigma^2)$ then

- ▶ $\bar{X} \sim N(\mu, \sigma^2/n)$
- ▶ $S^2 = \sum(X - \bar{X})^2/(n - 1)$ and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

- ▶ $\bar{X} \perp S^2$ and
- ▶
$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Simple Linear Regression

A simple linear regression model is obtained by estimating the intercept and slope in the equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \dots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$. The values of β_0, β_1 that minimize the sum of squares

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2,$$

are called the least squares estimators. They are given by:

- ▶ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- ▶ $\hat{\beta}_1 = r \frac{S_y}{S_x}$

r is the correlation between y and x , and S_x, S_y are the sample standard deviations of x and y respectively.

Case Study 1: The Spock Conspiracy Trial

- ▶ Boston, 1968
- ▶ Dr. Benjamin Spock (paediatrician and author) on trial for conspiring to violate the Selective Service Act.
- ▶ Accused of encouraging people to dodge military draft by his books that advised on how mothers should raise children.
- ▶ Spock's jury had NO women.

Judge

Q: Is there evidence of gender bias in the jury selection for Spock's trial?

Case Study 1: Jury selection

- ▶ 300 names selected at random from city directory
- ▶ 35 to 200 jurors randomly selected (this group is called the venire)
- ▶ Then non-random selection or exclusion of jurors from the venire by both defence and prosecution
- ▶ For Spock's trial, only 1 woman in the venire but she was then dismissed by prosecution
- ▶ Defence argued that Spock's judge had history of women being underrepresented on his venires.

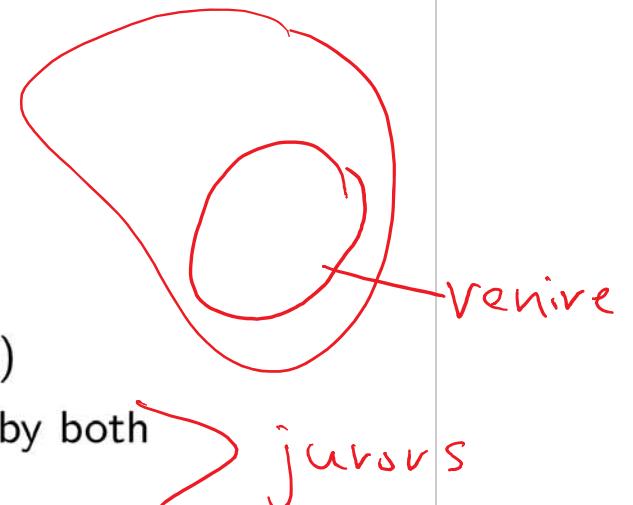
► Compared composition of recent venires of 6 other judges with that of Spock's judge

► Data: percent of women in each venire

Cat

JUDGE
S
S
S
O
O

% of women — Numeric.
0 - 100
0 - 100
|
|
|



Case Study 1: Two Key Questions

S vs Others

S	O
---	---

- ▶ Q1. Is there evidence that women are underrepresented on Spock's judge's venires when compared to other judges?
- ▶ Q2. Is there evidence that there are differences in women's representation in venires of the other 6 judges?
- ▶ Q: Conduct the relevant hypothesis test to answer Q1. Include the necessary assumptions, justifications and elements of a hypothesis test. What is your conclusion in plain English?

Among others

T	I	I	I	-	I	I	6
---	---	---	---	---	---	---	---

Case Study 1: The Spock Conspiracy Trial Data

The data is shown below.

```
#Juries data  
juries<-read.csv(  
  "/Users/Shivon/STA303_1002/LectureNotes/Lec1/juries.csv", header=T)
```

attach(juries)

#head(juries) — 1st 6 rows of data

PERCENT (of women in venires by judges)

```
## [1] 6.4 8.7 13.3 13.6 15.0 15.2 17.7 18.6 23.1 16.8 30.8 33.6 40.  
## [15] 27.0 28.9 32.0 32.7 35.5 45.6 21.0 23.4 27.5 27.5 30.5 31.9 32.  
## [29] 33.8 24.3 29.7 17.7 19.7 21.5 27.9 34.8 40.2 16.5 20.7 23.5 26.  
## [43] 29.5 29.8 31.9 36.2
```

JUDGE

$n=46$

```
## [1] SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS  
## [11] A A A A B B B B B  
## [21] C C C C C C C C C  
## [31] D E E E E E E F F  
## [41] F F F F F F  
## Levels: A B C D E F SPOCKS
```

\bar{x} | 10

N_1
 N_2

—
—

—

—

—
—
—
—

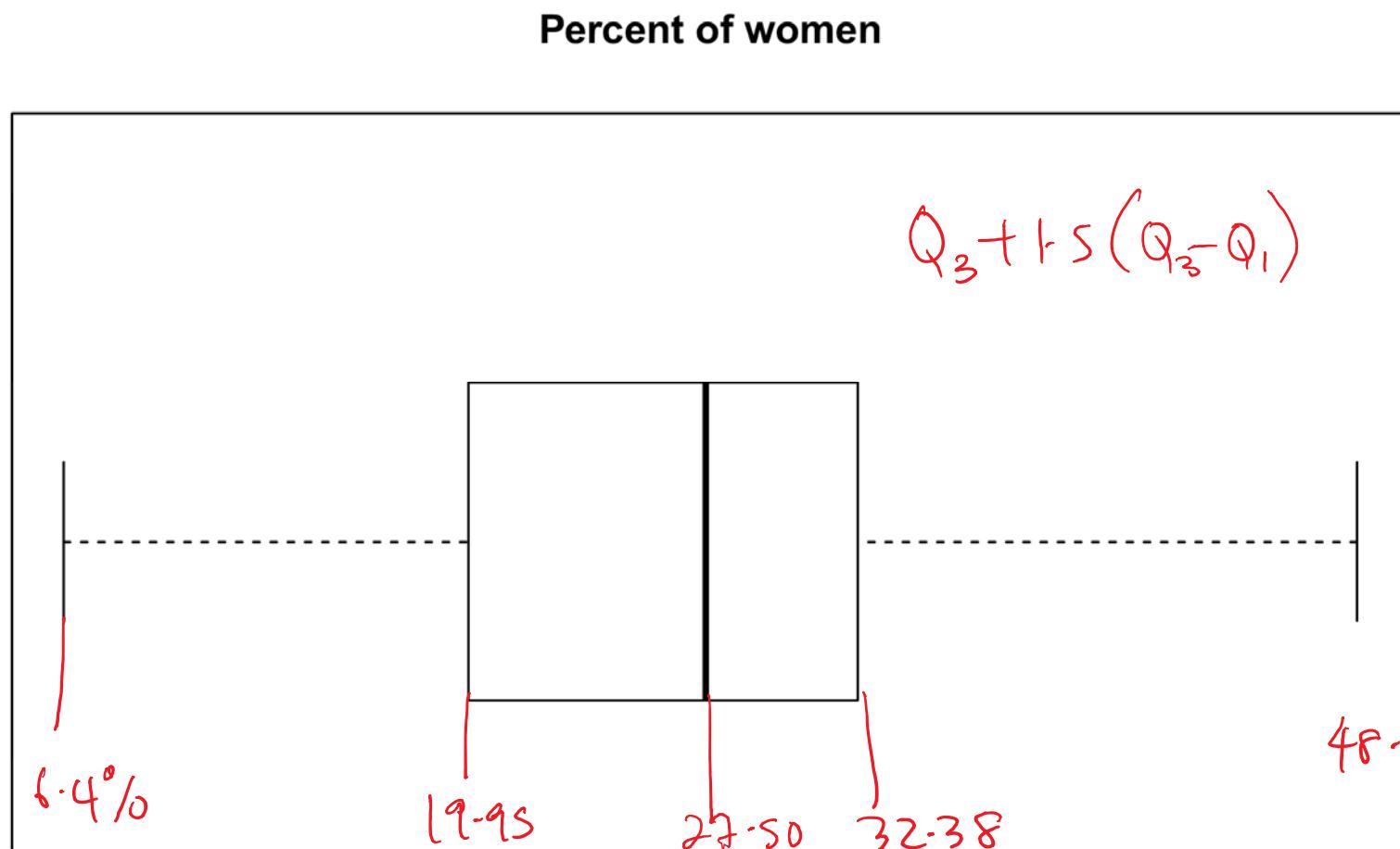
$\sqrt{46}$

Case Study 1: Data summary

```
summary(PERCENT)
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
## 6.40 19.95 27.50 26.58 32.38 48.90
```

```
boxplot(PERCENT, horizontal=T, main="Percent of women")
```



upper fence

Outlier if $> Q_3 + 1.5(Q_3 - Q_1)$

or $< Q_1 - 1.5(Q_3 - Q_1)$

$h = 46$.

lower fence.

Mean < Median

1.5 IQR Rule

for Outlier

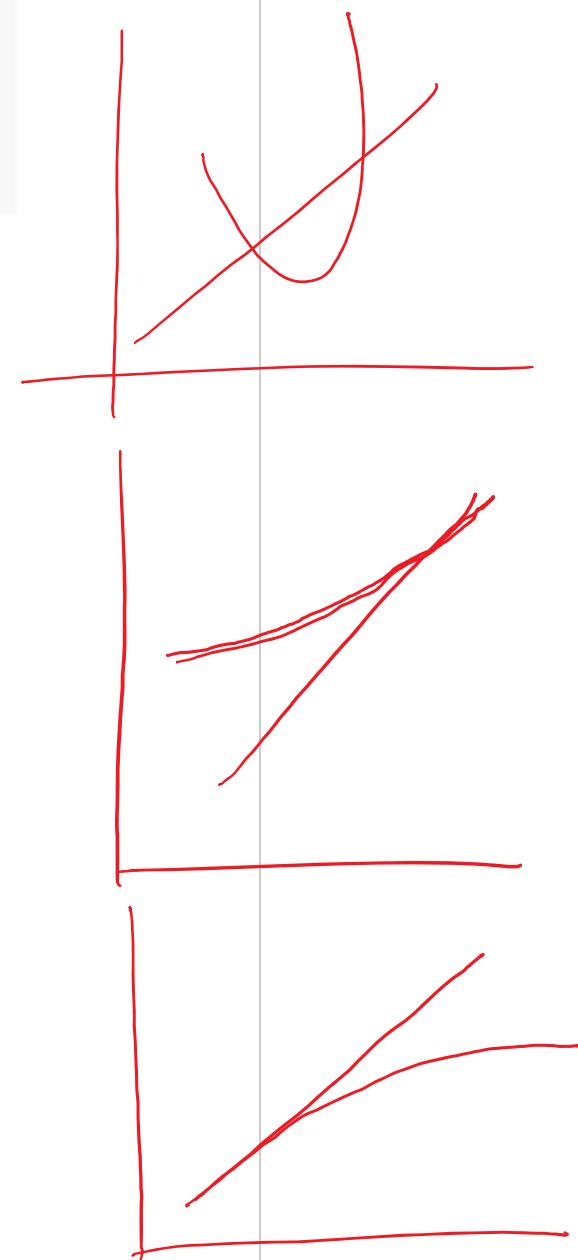
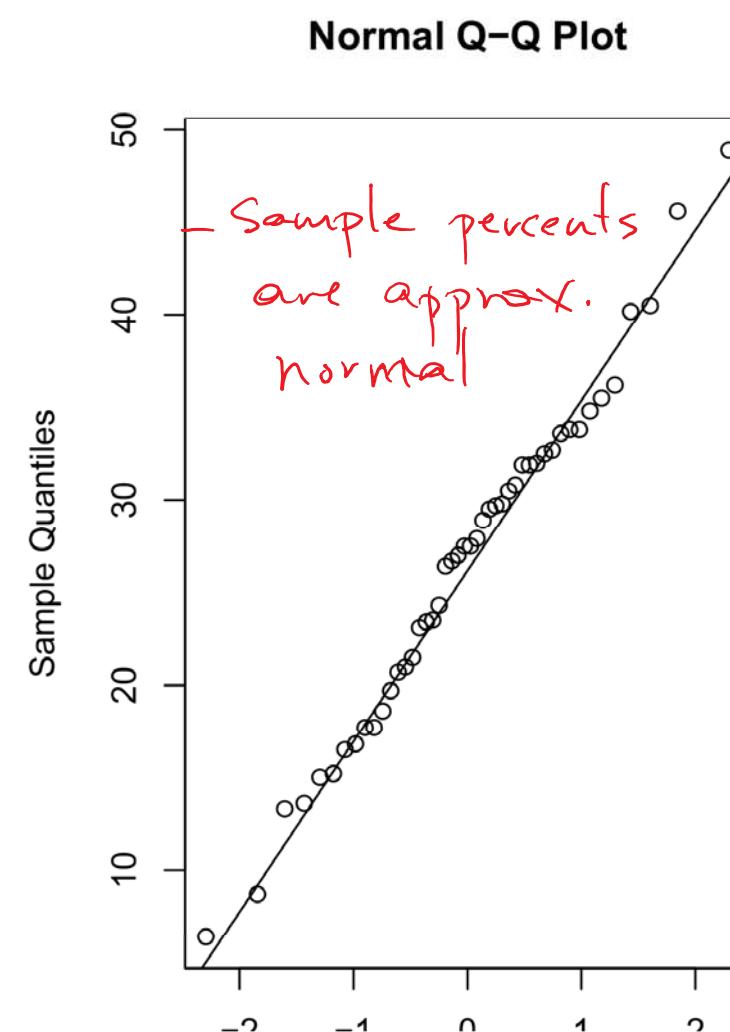
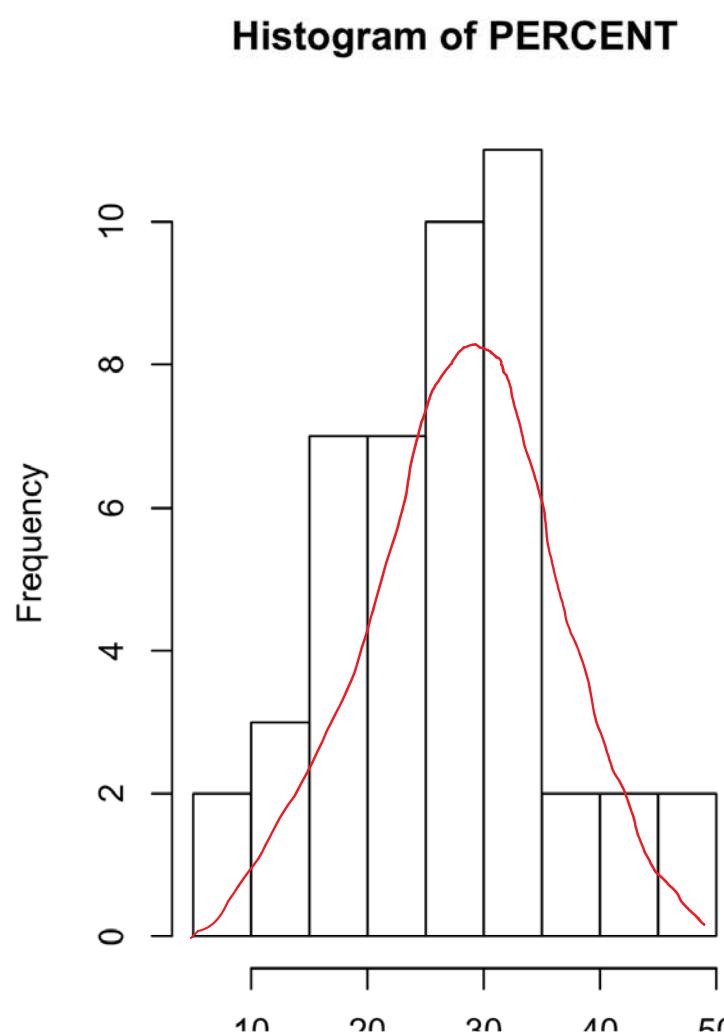
detection

(default in
boxplot())

48.9% help(boxplot)

Case Study 1: Check Normality

```
par(mfrow=c(1,2))
hist(PERCENT)
qqnorm(PERCENT)
qqline(PERCENT)
```



Case Study 1: Check Normality

```
shapiro.test(PERCENT)

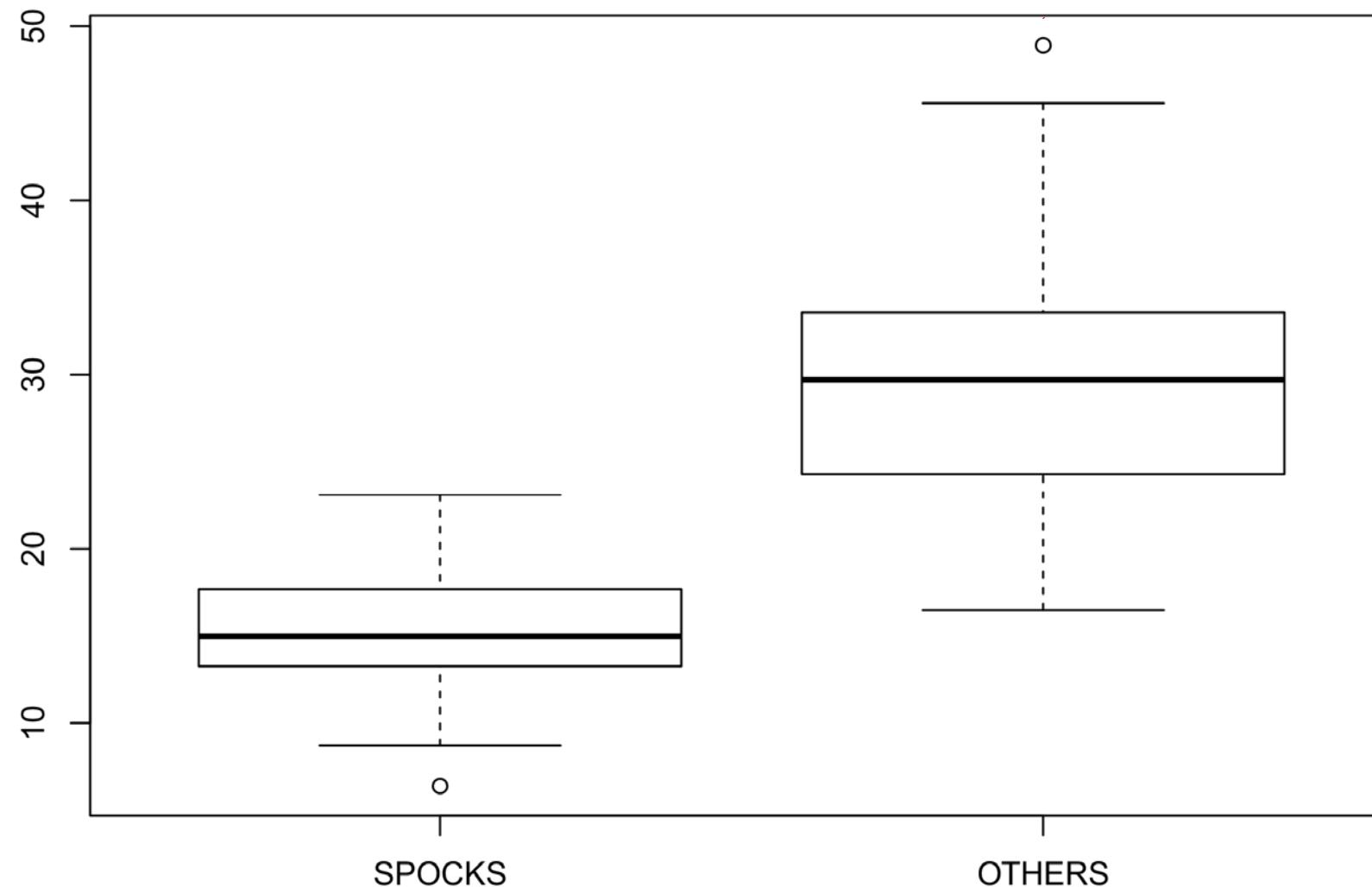
##  
## Shapiro-Wilk normality test  
##  
## data: PERCENT
## W = 0.98763, p-value = 0.9013
```

② Test statistic

- ① $n=46$.
- H_0 : Data are Normal
- H_a : Data are NOT Normal
- ③ large p-value
- ④ Do not reject H_0
— Evidence that data are normal.

Case Study 1: Two Sample t-tests

```
groupS<-PERCENT [JUDGE=="SPOCKS"]  
groupNS<-PERCENT [JUDGE!="SPOCKS"]  
boxplot(groupS, groupNS,xlab="JUDGE",names=c("SPOCKS", "OTHERS"))
```



1.5 IQR Rule

Two-sample t-tests

- Purpose: To compare two population means
- Data: Two random samples X_1, \dots, X_{n_x} and Y_1, \dots, Y_{n_y} of sizes n_x and n_y from population 1 and population 2
- Null Hypothesis:

$$H_0 : \mu_x - \mu_y = D_0 \text{ (typically } D_0 = 0)$$

Assumptions:

- The two samples are iid from approximately Normal populations.
- The two samples are independent of each other.
- Test statistic:

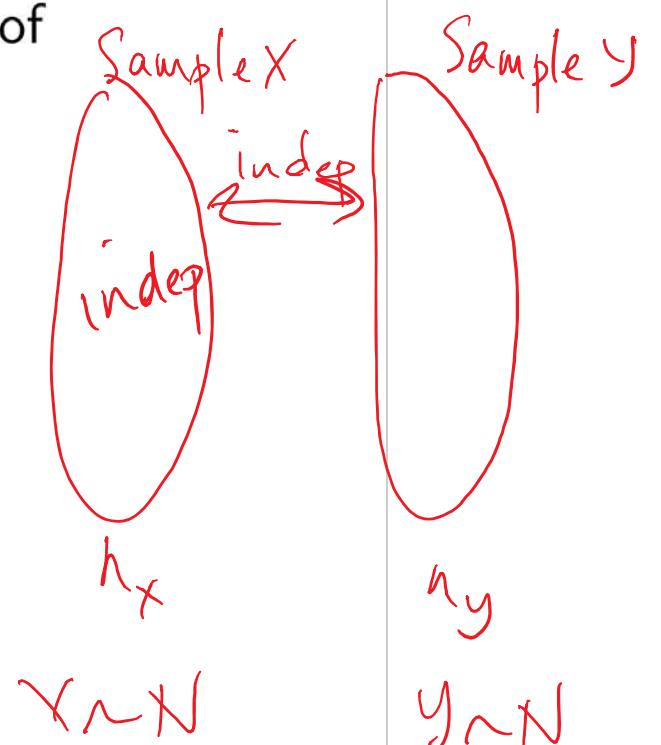
$$t = \frac{(\bar{x} - \bar{y}) - D_0}{se(\bar{x} - \bar{y})}$$

Q: How do we estimate this standard error ("se") - standard deviation of $\bar{x} - \bar{y}$?

Intro to 1-way ANOVA

$$\begin{aligned} se(\bar{x} - \bar{y}) &= \sqrt{\text{Var}(\bar{x} - \bar{y})} \\ &= \sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y} \end{aligned}$$

$$H_0: \mu_x = \mu_y$$



$$\begin{aligned} \text{Var}(\bar{x} - \bar{y}) &= \text{Var}(\bar{x}) + \text{Var}(\bar{y}) \\ &= \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} \end{aligned}$$

Case Study 1: Checking equal variance assumption

```
var(groupS)
```

```
## [1] 25.38945
```

```
var(groupNS)
```

```
## [1] 55.21632
```

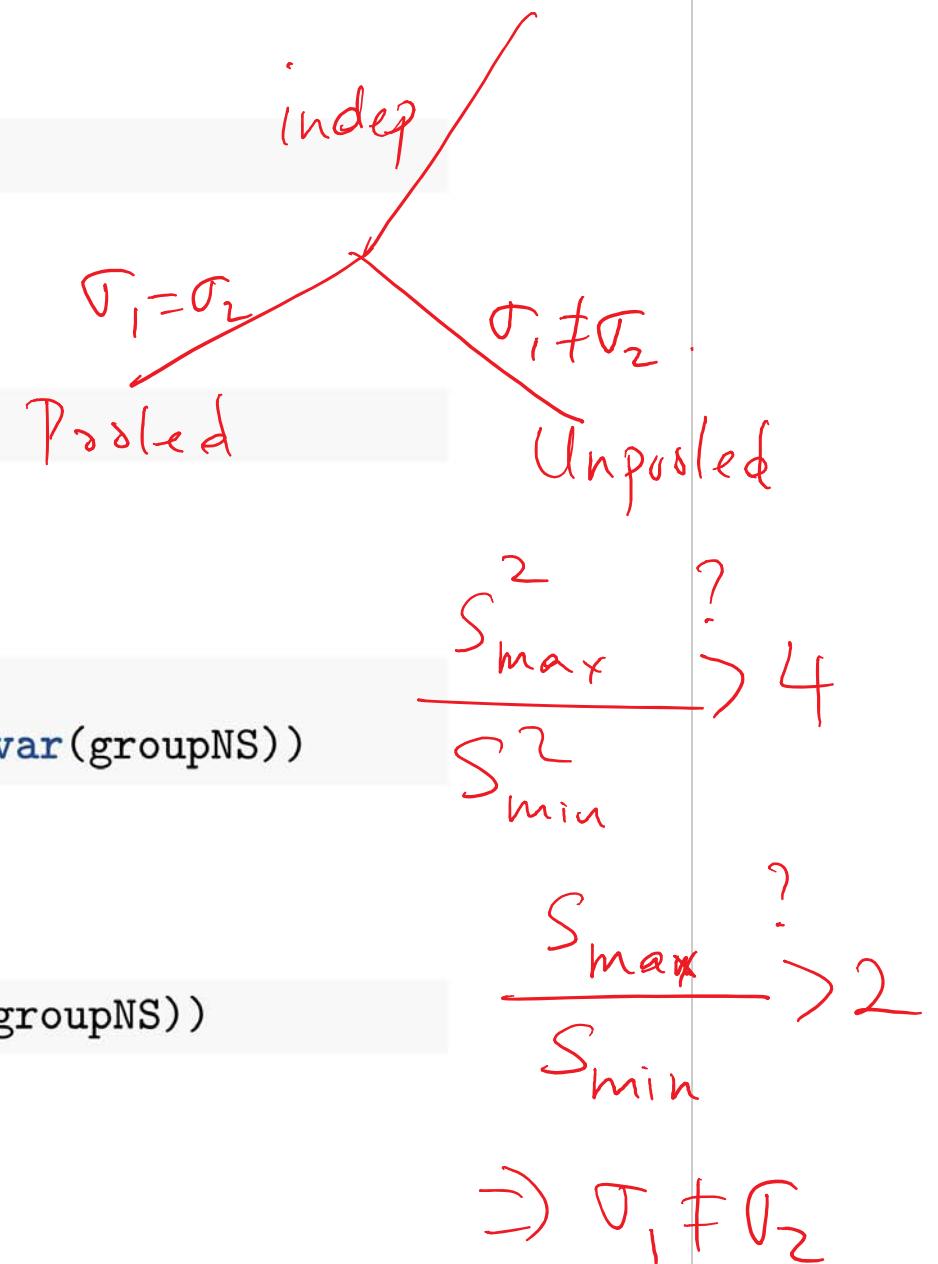
#Rule of Thumb

```
max(var(groupS), var(groupNS)) / min(var(groupS), var(groupNS))
```

```
## [1] 2.174775 < 4
```

```
max(sd(groupS), sd(groupNS)) / min(sd(groupS), sd(groupNS))
```

```
## [1] 1.474712 < 2
```



Rule of thumb for checking equal variances

- ▶ Test:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_a : \sigma_1^2 \neq \sigma_2^2$$

- ▶ Test statistic:

$$\frac{\text{larger sample variance}}{\text{smaller sample variance}} = \frac{S_{max}^2}{S_{min}^2}$$

- ▶ If test statistic is greater than 4, reject H_0

Variance Ratio F-test

- ▶ special case of Bartlett's test for homogeneity of variances (Bartlett, 1937)
- ▶ Null Hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$
- ▶ Underlying assumptions:
 - ▶ Random samples of sizes n_1 and n_2 are drawn from Normal populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively
 - ▶ Samples are independent
 - ▶ Samples are large (better when sample sizes are equal too)
- ▶ **Test statistic:**

$$F = \frac{S_1^2}{S_2^2} \sim_{H_0} F_{n_1-1, n_2-1}$$

- ▶ In R: var.test()
- ▶ For more than 2 variances:
 - ▶ bartlett.test()
 - ▶ Robust alternative: Levene's test (levene.test())

$$F = \frac{\frac{\chi^2_{n_1}}{n_1}}{\frac{\chi^2_{n_2}}{n_2}}$$

$$\frac{(n_1-1) S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$$

$$\frac{(n_2-1) S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

Case Study 1: Checking equal variance assumption

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

#F Test of Equal variances
`var.test(groupS, groupNS)`

```
##  

## F test to compare two variances  

##  

## data: groupS and groupNS  

## F = 0.45982, num df = 8, denom df = 36, p-value = 0.2482  

## alternative hypothesis: true ratio of variances is not equal to 1  

## 95 percent confidence interval:  

## (0.1789822, 1.7739665)  

## sample estimates:  

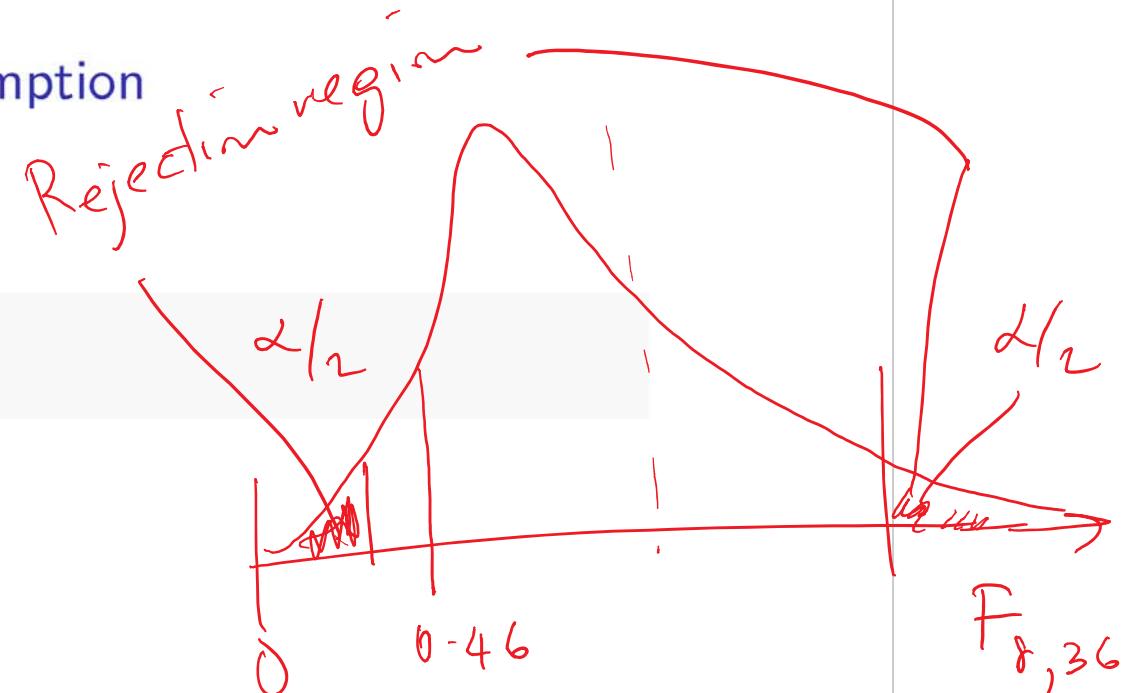
## ratio of variances  

## 0.4598178
```

⇒ No sig. results

$$(P=0.25) > (\alpha=0.10)$$

Evidence that variances are equal.



Concl: Sig. results if

$$P\text{-value} < \alpha$$

$$\alpha = P(\text{Type I Error})$$

$$\beta = P(\text{Type II Error})$$

≈ 0
 5%
 10%
 1%

Two-sample t-test (Satterthwaite approximation)

- ▶ Used when population variances cannot be assumed to be equal
- ▶ Test statistic: under H_0 ,

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \sim t_{\nu}$$

where

$$\nu = \frac{(s_x^2/n_x + s_y^2/n_y)^2}{\frac{(s_x^2/n_x)^2}{n_x-1} + \frac{(s_y^2/n_y)^2}{n_y-1}}$$

or $\min(n_x-1, n_y-1)$

- ▶ The df(degrees of freedom), ν is calculated by Satterthwaite approximation.
- ▶ ν may not be an integer so round down to the nearest integer

$$se(\bar{x} - \bar{y}) = \sqrt{\frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}}$$

Compare to $n_x + n_y - 2$

Pooled two-sample t-test

- ▶ Special case of two-sample t-test
- ▶ Assumes population variances are equal
- ▶ Pooled variance estimate

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

$$(\sigma_1^2 = \sigma_2^2)$$

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

- ▶ Test statistic: under H_0

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{s_p^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}} \sim t_{n_x + n_y - 2}$$

$$\sqrt{\frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}} \quad \sqrt{\sigma^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}$$

Case Study 1: Two sample (unpooled) t-tests

```
#Welch-Satterthwaite (Unpooled)  
t.test(groupS, groupNS, var.equal=F)
```

```
##  
## Welch Two Sample t-test  
##  
## data: groupS and groupNS  
## t = -7.1597, df = 17.608, p-value = 1.303e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -19.23999 -10.49935  
## sample estimates:  
## mean of x mean of y  
## 14.62222 29.49189
```

$\text{df} = 17$ ≈ 0

Case Study 1: Pooled t-test

```
#Pooled  
t.test(groupS, groupNS, var.equal=T)
```

```
##  
## Two Sample t-test  
##  
## data: groupS and groupNS  
## t = -5.6697, df = 44, p-value = 1.03e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -20.155294 -9.584045  
## sample estimates:  
## mean of x mean of y  
## 14.62222 29.49189
```

Spock's

others

$$9 + 37 - 2 = 44$$

⇒ Evidence that the
% of women are
different on venires
of Spock's judge vs
others.

Case Study 1: Paired t-test

```
#Paired  
t.test(groupS, groupNS, paired=TRUE)
```

Error in complete.cases(x, y): not all arguments have the same length

- Need equal sample sizes
- Need paired (dependent samples) data.

Case Study 1: Pooled t-test (Left tailed)

```
#Left-tailed Pooled  
t.test(groupS,groupNS,alternative="less",var.equal=TRUE)  
  
##  
## Two Sample t-test  
##  
## data: groupS and groupNS  
## t = -5.6697, df = 44, p-value = 5.148e-07  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## [-Inf -10.463]  
## sample estimates:  
## mean of x mean of y  
## 14.62222 29.49189
```

⇒ Evidence of lower %
of women on review
of Spock's Judge vs
others.

Simple Linear Model Approach (Dummy variable)

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

$$X_i = \mathbb{1}_{A,i} = \begin{cases} 1 & \text{if } i\text{th observation is from "group A"} \\ 0 & \text{if } i\text{th observation is NOT from "group A"} \end{cases}$$

Assumptions:

- ▶ The linear model is appropriate
- ▶ Gauss-Markov properties:
 - ▶ $E(\epsilon_i) = 0$
 - ▶ $\text{Var}(\epsilon_i) = \sigma^2$: Uncorrelated errors
- ▶ $\epsilon_i \sim \text{Normal}$

$$\epsilon_i \sim N(0, \sigma^2)$$

Residual
plots.

Intro to 1-way ANOVA

Simple Linear Model: The Hypothesis Test

Test:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 \neq 0$$

$$H_0: \mu_s = \mu_o$$

- The slope, β_1 , captures the difference in means between groups
- Proof:

- $E(Y|A) = E(Y|X == 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$
- $E(Y|A^c) = E(Y|X == 0) = \beta_0 + \beta_1 \times 0 = \beta_0$

► Hence,

$$\beta_1 = E(Y|A) - E(Y|A^c) = E(Y|X == 1) - E(Y|X == 0)$$

Test statistic: Under the assumptions and H_0 ,

$$t = \frac{b_1}{se(b_1)} \sim t_{N-2=n_A+n_{others}-2}$$

$$M_A \quad M_{A^c}$$

Pooled t

$$\hat{\beta}_0 = \bar{y}_{A^c}$$

$$\hat{\beta}_1 = \bar{y}_A - \bar{y}_{A^c}$$

Case Study 1: Simple Linear Regression Approach

```
X=c(rep(1,length(groupS)), rep(0,length(groupNS))) #X==1-Spock's judge,  
Y=PERCENT; model1<-lm(Y~X); summary(model1)
```

```
##  
## Call:  
## lm(formula = Y ~ X)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -12.9919  -4.6669   0.2581   3.7854  19.4081  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  29.492     1.160   25.42 < 2e-16 ***  
## X          -14.870     2.623  -5.67 1.03e-06 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 7.056 on 44 degrees of freedom  
## Multiple R-squared:  0.4222, Adjusted R-squared:  0.409  
## F-statistic: 32.15 on 1 and 44 DF, p-value: 1.03e-06
```

$$T^2_{44} = F_{1,44}$$
$$(-5.67)^2 = 32.15$$

$$H_0: \beta_1 = 0$$

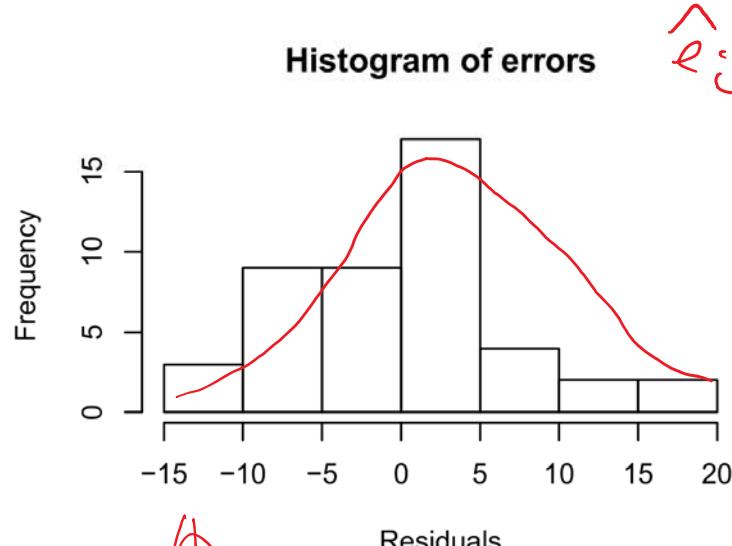
$$R^2$$

Case Study 1: Regression diagnostics

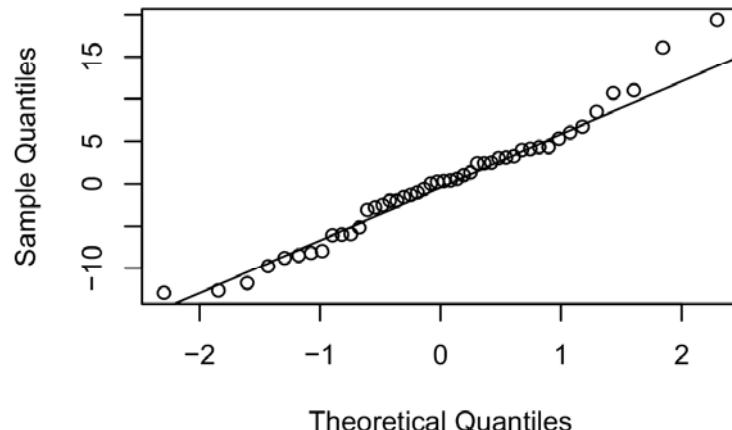
```
yhats=fitted(model1)
errors=residuals(model1)
# par(mfrow=c(2,2)) #partition plot window
# #plot (1,1)- histogram of residuals
# hist(errors, xlab="Residuals", breaks=5)
# #plot(1,2)- residuals vs index(time) with zero line
# # plot(errors)
# abline(0,0)
# #plot(2,1)-normal qq plot of residuals with qqline
# qqnorm(errors)
# qqline(errors)
# #plot(2,2)-residuals vs fitted values with zero line
# plot(yhats, errors, xlab="Fitted values", ylab="Residuals")
# abline(0,0)
```

Case Study 1: Regression diagnostics

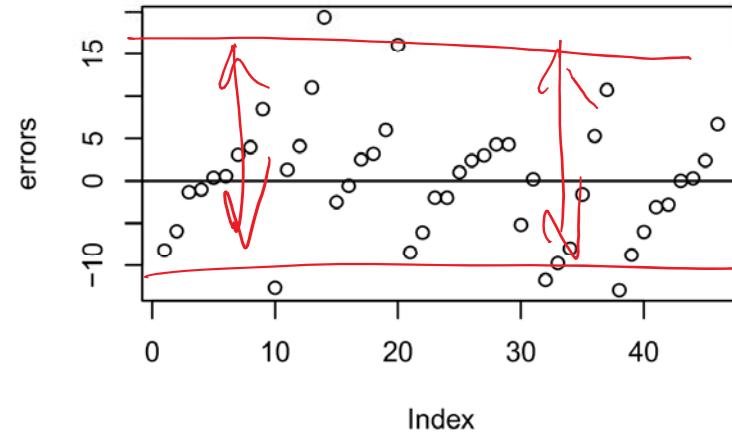
Histogram of errors



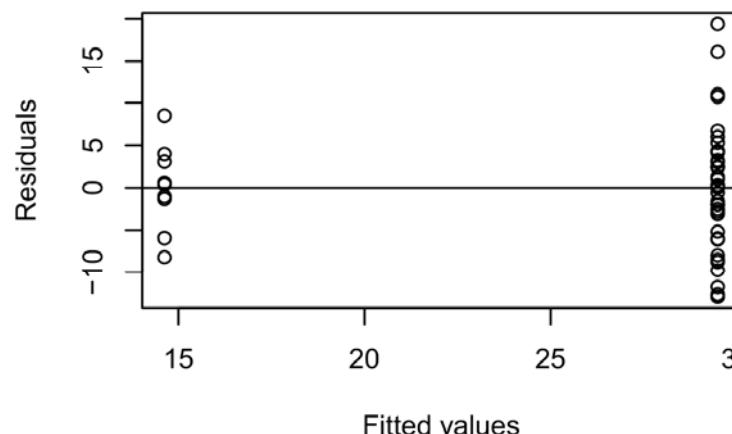
Normal Q-Q Plot



$$e_i \sim N(0, \sigma^2)$$



Constant variation



Normality satisfied

Case Study 1: One-way ANOVA approach

```
#ANOVA approach  
anova(model1)
```

```
## Analysis of Variance Table  
##  
## Response: Y  
##           Df Sum Sq Mean Sq F value    Pr(>F)  
## X          1 1600.6 1600.62 32.145 1.03e-06 ***  
## Residuals 44 2190.9   49.79  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H_0: \mu_A = \mu_{A^C}$$

Case Study 1: Partial results for (Q1)

Sample	SPOCK'S	OTHER
Mean	14.6222	29.4919
Standard deviation	5.0388	7.4308
Sample size	9	37

Hypothesis Test	Partial results
Equal variances assumed	Yes
t-test statistic	-5.67
<i>df</i>	44
P-value	≈ 0
Conclusion	Reject H_0

Notes:

- ▶ Equivalence: Pooled 2-sample t is a special case of One-way ANOVA
- ▶ Diagnostics: Gauss-Markov assumptions satisfied
- ▶ Caution: Unequal sample sizes

Robustness of t

- ▶ t-procedures are robust against assumptions of normality.
- ▶ In other words, t-procedures are often valid even when the assumption of normality is violated.
- ▶ They are not robust against strong skewness or outliers
- ▶ Can be used when sample size is small

- ▶ Non-parametric tests or “Distribution free” tests do not require that data follow any specific distribution.

Non-parametric alternatives

Gaussian	“Distribution free”
1-sample t	Sign test, Wilcoxon signed-rank test
2-sample t	Wilcoxon rank-sum test

In R: See wilcox.test()

R functions used

```
summary()  
plot()  
boxplot()  
t.test()  
    pnorm()  
    qqnorm()  
    qqline()  
shapiro.test()  
var.test()  
    lm()  
    fitted()  
    residuals()  
anova()
```

STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



January 16-18, 2018

One-way ANOVA

STA 303/1002: Week 2 Outline

- ▶ The General Linear Model
- ▶ One-way ANOVA
 - ▶ With $G=2$
 - ▶ With $G > 2$
- ▶ Case Study 1 continued
- ▶ Diagnostics- checking model assumptions
 - ▶ Normality of errors
 - ▶ Constant variance
 - ▶ Uncorrelated errors
- ▶ Multiple comparisons: Bonferroni and Tukey's

One-way ANOVA

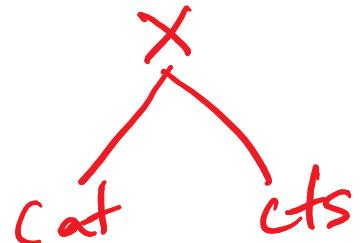
Week 1 Review

- ▶ Review:

- ▶ One sample t-test $H_0: \mu = \mu_0$
- ▶ Two sample t-tests (`t.test()` or `summary(lm())` or `anova()` in R)
- ▶ Testing equal variances
- ▶ Assessing normality
- ▶ Case Study 1: Question 1

One-way ANOVA

The General Linear Model



y -cts
 x -cts

- ▶ Response, Y is continuous
- ▶ Explanatory variable(s), X is(are) categorical and/or continuous
- ▶ Y is linear in the β 's-i.e. no predictor is a linear function or combination of other predictors
- ▶ In R: `lm()`

Review of Regression in Matrix Terms

Model:

$$Y_i = \boxed{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}} + \epsilon_i$$

for $i = 1, \dots, N$

Matrix Form: $Y = X\beta + \epsilon$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix}_{N \times (p+1)}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1) \times 1}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}_{N \times 1}$$

One-way ANOVA

Least-squares Estimates for β

- ▶ $\hat{\beta} = (X'X)^{-1}X'Y$
- ▶ $X'X$ has dimension $(p + 1) \times (p + 1)$
- ▶ Need $X'X$ to be of full rank to be invertible:
 - ▶ $\text{rank}(X'X) = \text{rank}(X)$
 - ▶ Need X to be rank $p + 1$
 - ◀▶ The columns of X must be linearly independent ▶

Gen LM: Hypothesis and Assumptions

- ▶ Null Hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$G=2$

(GLM) $H_0: \beta_1 = 0$

- ▶ Assumptions:

- ▶ Linear Model is appropriate: Errors have zero expectation,
 $E[\epsilon_i] = 0$

- ▶ Homoscedasticity of variances: Errors have constant variance,
 $Var(\epsilon_i) = \sigma^2$

- ▶ Errors are uncorrelated

- ▶ Errors are jointly normally distributed

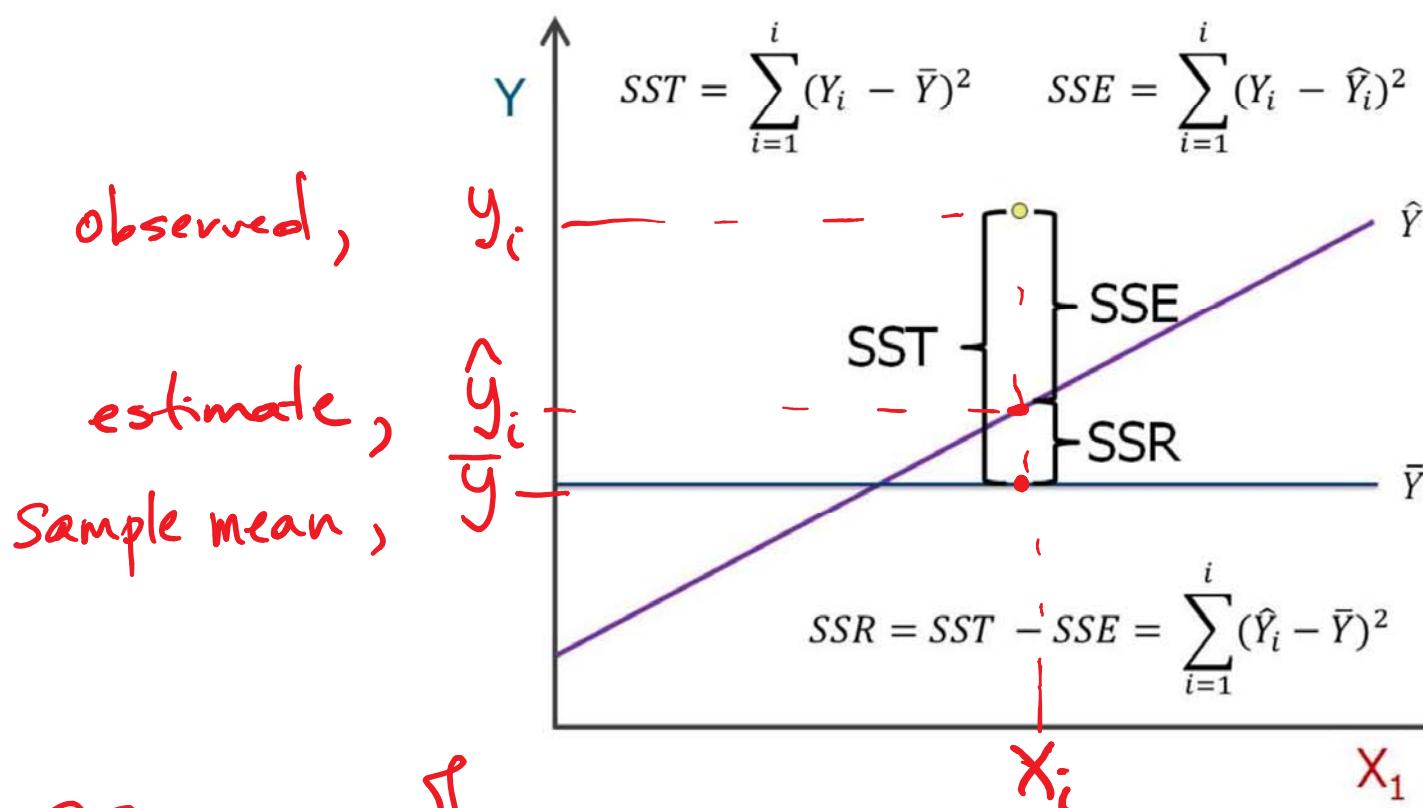
Indep.
observations

One-way
ANOVA

$$H_1: M_A \neq M_{AC}$$

GLM: Sum of Squares Decomposition

Aim, $E(Y_i) = \mu_y$



SS - sum of squares

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}), \quad \boxed{\hat{Y}_i = X_i \beta}$$

One-way ANOVA

SS Total

SSE = RSS

SS Reg

$$\hat{e}[y_i | x_i] = \hat{y}_i$$

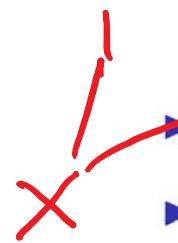
$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N}$$

$$\hat{y}_i = X_i \hat{\beta}$$

(model)

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

One-Way ANOVA



- Response/Outcome is continuous $\rightarrow Y$
- One factor (categorical/grouping variable) with at least 2 levels ($G \geq 2$)
- Aim: Compare G group means

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_G = \mu, \quad H_a : \exists i \neq j \text{ s.t. } \mu_i \neq \mu_j$$

- Predictors are indicator variables that classify the observations one way (into G groups)
 - Special case of a general linear model (GLM)
 - Equivalent to GLM with one-way classification (one factor)
 - GLM uses $\underline{G - 1}$ dummy variables.
- ANOVA: compare means by analyzing variability

(Q1)

One-way ANOVA

$$G=2$$

$$y_i = \beta_0 + \beta_1 X_{iA}$$

$$\text{where } X_{iA} = \begin{cases} 1 & \text{if } i\text{th obs in} \\ 0 & \text{o-w group}\end{cases}$$

Eg

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$$\Rightarrow X_{i1} + X_{i2} = 1$$

lin. dep. cols.

Brief History of ANOVA

- ▶ Dates back to early work by R. A. Fisher in 1918 on mathematical genetics
- ▶ Further developed by Fisher in 1920
- ▶ The convenient acronym - ANOVA was coined much later by John W. Tukey (1915-2000), the pioneer of exploratory data analysis (EDA)
- ▶ The test developed was named the F in his honour

Data layout and Notation

Treatment or factor levels

	1	2	...	G
1	Y_{11}	Y_{21}	...	Y_{G1}
2	Y_{12}	Y_{22}	...	Y_{G2}
	:	:		:
	Y_{1n_1}	Y_{2n_2}	...	Y_{Gn_G}

Group means

	Sample mean	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$...	$\bar{Y}_{G.}$	$\bar{Y}_{..}$

Group

	Sample variance	S_1^2	S_2^2	...	S_G^2

$$\bar{Y}_{g.} = \frac{\sum_{j=1}^{n_g} Y_{gj}}{n_g}$$

One-way ANOVA

$n_g = \text{group sizes}$

$$\bar{Y}_{..} = \frac{\sum_{i=1}^N Y_i}{N}$$

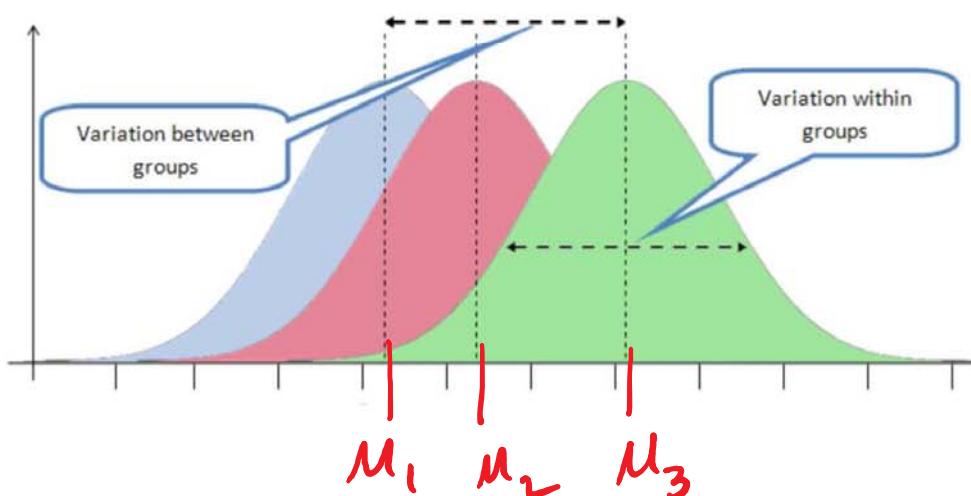
$$S_g^2 = \frac{1}{n_g - 1} \sum_{j=1}^{n_g} (Y_{gj} - \bar{Y}_{g.})^2$$

One-way ANOVA Assumptions

- ▶ The G samples are independently drawn from G specific populations with unknown means $\mu_1, \mu_2, \dots, \mu_G$.
- ▶ Each population is normally distributed.
- ▶ Each population has the same variance, σ^2 .

Compactly written:

$$E_i \sim \text{Normal}(\mathbf{0}_G, \sigma^2 \mathbf{I})$$



One-way ANOVA

One-way: Expectations and Estimates

- Expected values of Y , $\cancel{\mu}_i = E(Y_i)$
- Predicted values of Y , \hat{Y}_i
- Estimates of coefficients, $\hat{\beta}$

Parameter

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{G-1} x_{i,G-1} + \epsilon_i$$

$$E(y_i | x_{i1} = 1) = \beta_0 + \beta_1$$

$$E(Y_i) = \begin{cases} \beta_0 + \beta_1 \\ \vdots \\ \beta_0 + \beta_{G-1} \\ \beta_0 \end{cases}, \hat{Y}_i = \begin{cases} b_0 + b_1 \\ \vdots \\ b_0 + b_{G-1} \\ b_0 \end{cases}, \hat{\beta} = \begin{cases} b_0 = \bar{y}_G \\ b_1 = \bar{y}_1 - \bar{y}_G \\ \vdots \\ b_{G-1} = \bar{y}_{G-1} - \bar{y}_G \end{cases}$$

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}$$

$$H_0: \beta_1 = 0$$

$b_1 = \bar{y}_A - \bar{y}_{A^c}$

$N_A = N_{A^c}$

One-way ANOVA

by I.S. estimation
 $(G=2, \hat{\beta}_0, \hat{\beta}_1)$

$$\begin{aligned} \sum_{i=1}^N x_{ii} &= n_i \\ \sum_{i=1}^N x_{ii}^2 &= n_i \end{aligned}$$

Decomposition of SST

$$\begin{aligned} SST &= \sum_i^N (Y_i - \bar{Y})^2 = SS_{Reg} + RSS \\ &= \sum_i^N (\hat{Y}_i - \bar{Y})^2 + \sum_i^N (Y_i - \hat{Y}_i)^2 \end{aligned}$$

Model *Error*

- ▶ $N = n_1 + \dots + n_G$
- ▶ \hat{Y}_i = mean of observations for group g from which the i th observation belongs
- ▶ \hat{Y}_i is one of $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_G$
- ▶ $\bar{Y} = \bar{Y}_{..}$ is the grand mean

One-way ANOVA Table

$\cancel{SS/DF} \approx \text{Variances}$.

SOURCE	DF	SS	MS	F
Model	G-1	SSReg	$\text{MSReg} = \text{SSReg}/G-1$	MSReg/MSE
Error	N-G	RSS	$\text{MSE} = \text{RSS}/N-G$	
TOTAL	N-1	SST		

- ▶ SSReg: “between groups” SS
- ▶ RSS: “within groups” SS
- ▶ Overall idea: If between groups SS is larger than within groups SS, there is evidence that means are different

Which group means differ? Which is bigger- SSReg or RSS?

within

SSReg < RSS

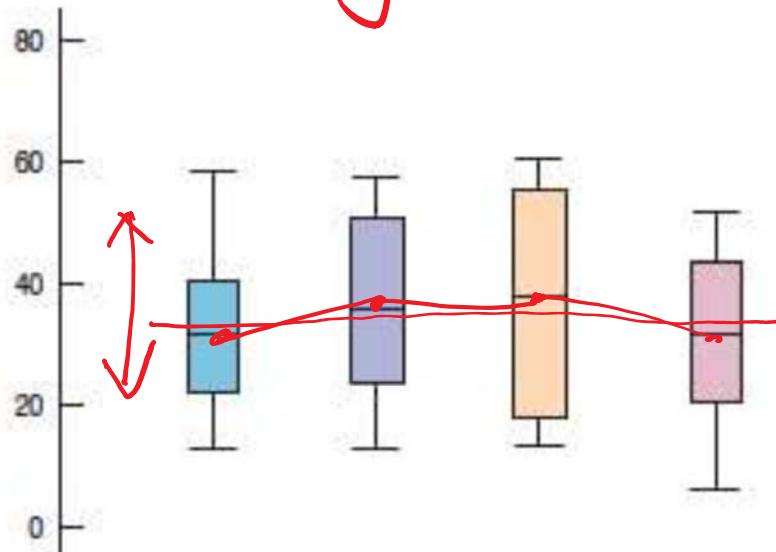


Figure 25.2

It's hard to see the difference in the means in these boxplots because the spreads are large relative to the differences in the means.

(*SDM, 2nd Canadian ed. by De Veaux et. al.*)

SSReg > RSS

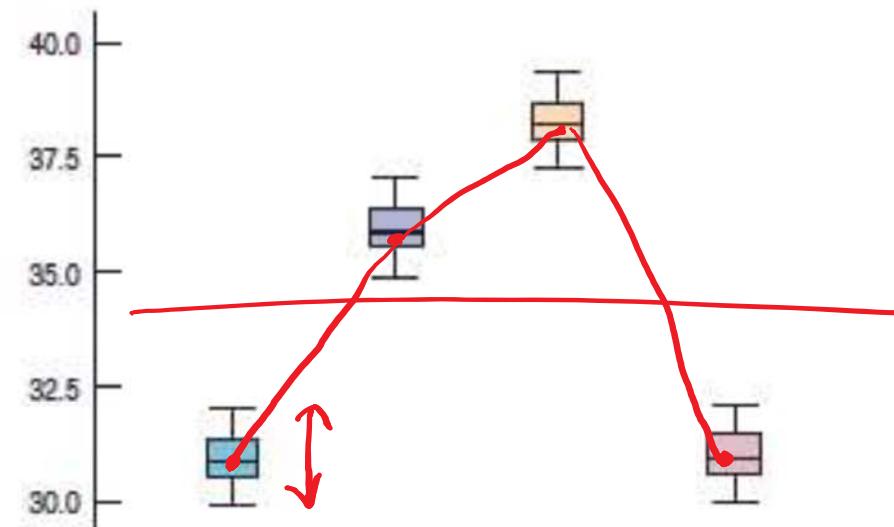


Figure 25.3

In contrast with Figure 25.2, the smaller variation makes it much easier to see the differences among the group means.

3rd .

One-way ANOVA

Derivation of SS's: SSReg and RSS

$$\begin{aligned}SS_{reg} &= \sum_i^N (\hat{Y}_i - \bar{Y})^2 \\&= \sum_g^G n_g (\bar{Y}_g - \bar{Y})^2\end{aligned}$$

$$\begin{aligned}RSS &= \sum_i^N (Y_i - \hat{Y}_i)^2 \\&= \sum_{g=1}^G \sum_{(g)} (Y_i - \bar{Y}_g)^2\end{aligned}$$

- ▶ g is the group index
- ▶ \hat{Y}_i is one of $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_G$
- ▶ $\sum_{(g)}$ -summation over observations in group g

Case Study 1 continued: The Spock Conspiracy Trial

One-way ANOVA

Case Study 1: The Spock Conspiracy Trial

Recall the 2 main questions:

(Q1) Is there evidence that women are underrepresented on Spock's judge's venires when compared to other judges?

(Q2) Is there a difference among the 6 other judges?

(A1): Two-sample t-test/ Simple linear regression model with 1 dummy predictor variable/ One-way ANOVA with G=2

(A2): Multiple linear regression model with 5 dummy predictor variables/ One-way ANOVA with G=6

Overall task: Compare the percent of women on venires of all 7 judges

One-way ANOVA

Case Study 1: The Spock Conspiracy Trial Data

Get the data (from desktop):

```
#Juries data
juries<-read.csv(
  "/Users/Shivon/STA303_1002/LectureNotes/Lec1/juries.csv", header=T)
attach(juries)
head(juries)

##    PERCENT   JUDGE
## 1      6.4 SPOCKS
## 2      8.7 SPOCKS
## 3     13.3 SPOCKS
## 4     13.6 SPOCKS
## 5     15.0 SPOCKS
## 6     15.2 SPOCKS
```

Case Study 1: The Spock Conspiracy Trial Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case0502
library(Sleuth3)
#Juries data
jury = case0502
attach(jury)
head(jury)
```

```
##   Percent    Judge
## 1     6.4 Spock's
## 2     8.7 Spock's
## 3    13.3 Spock's
## 4    13.6 Spock's
## 5    15.0 Spock's
## 6    15.2 Spock's
```

Ramsey & Schafer
The Statistical Sleuth
3rd ed.

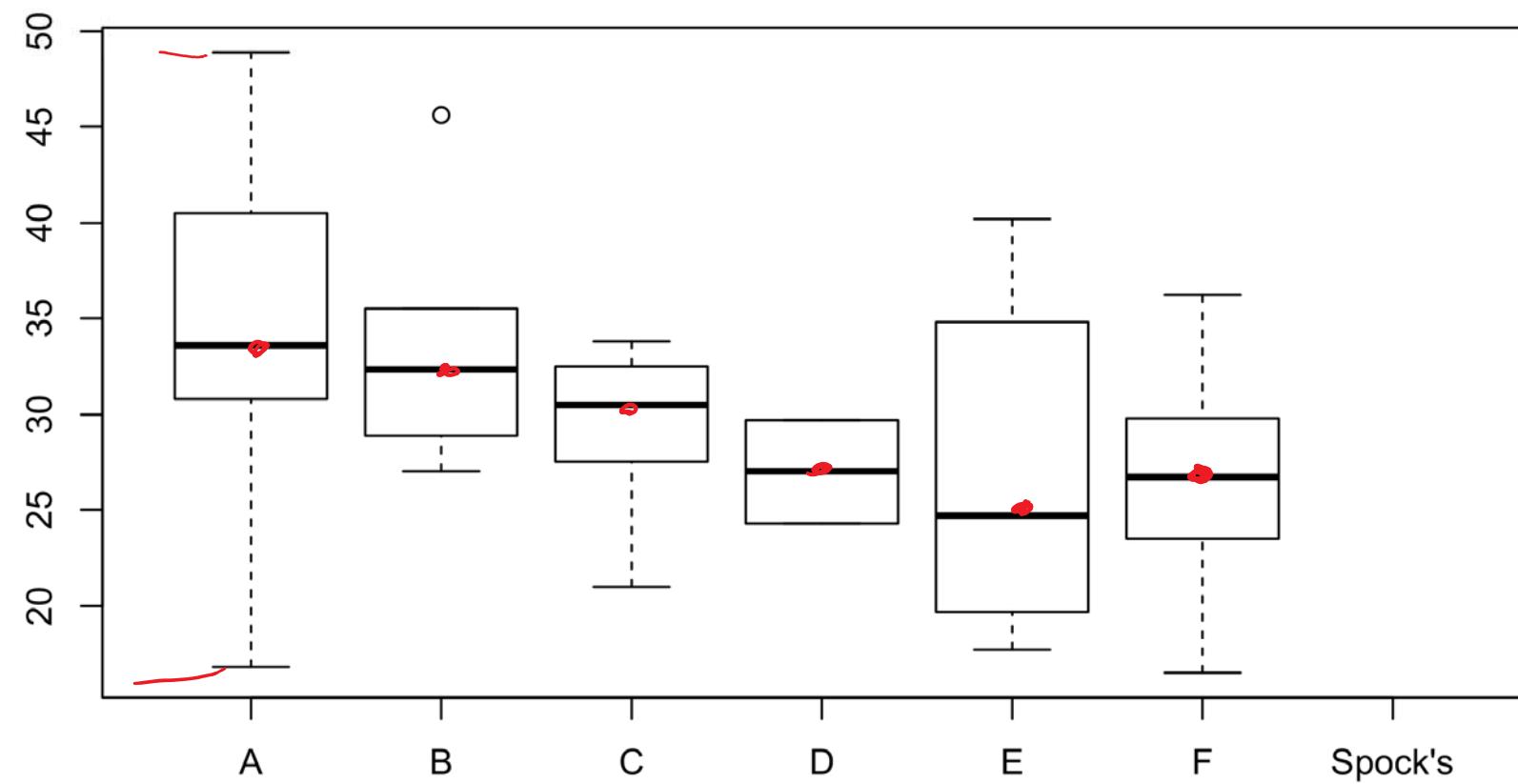
Case Study 1: How many venires for each Judge?

```
table(Judge)
## Judge
##      A      B      C      D      E      F Spock's
##      5      6      9      2      6      9      9

with(jury, tapply(Percent, Judge, mean))
##          A          B          C          D          E          F Spock's
## 34.12000 33.61667 29.10000 27.00000 26.96667 26.80000 14.62222
```

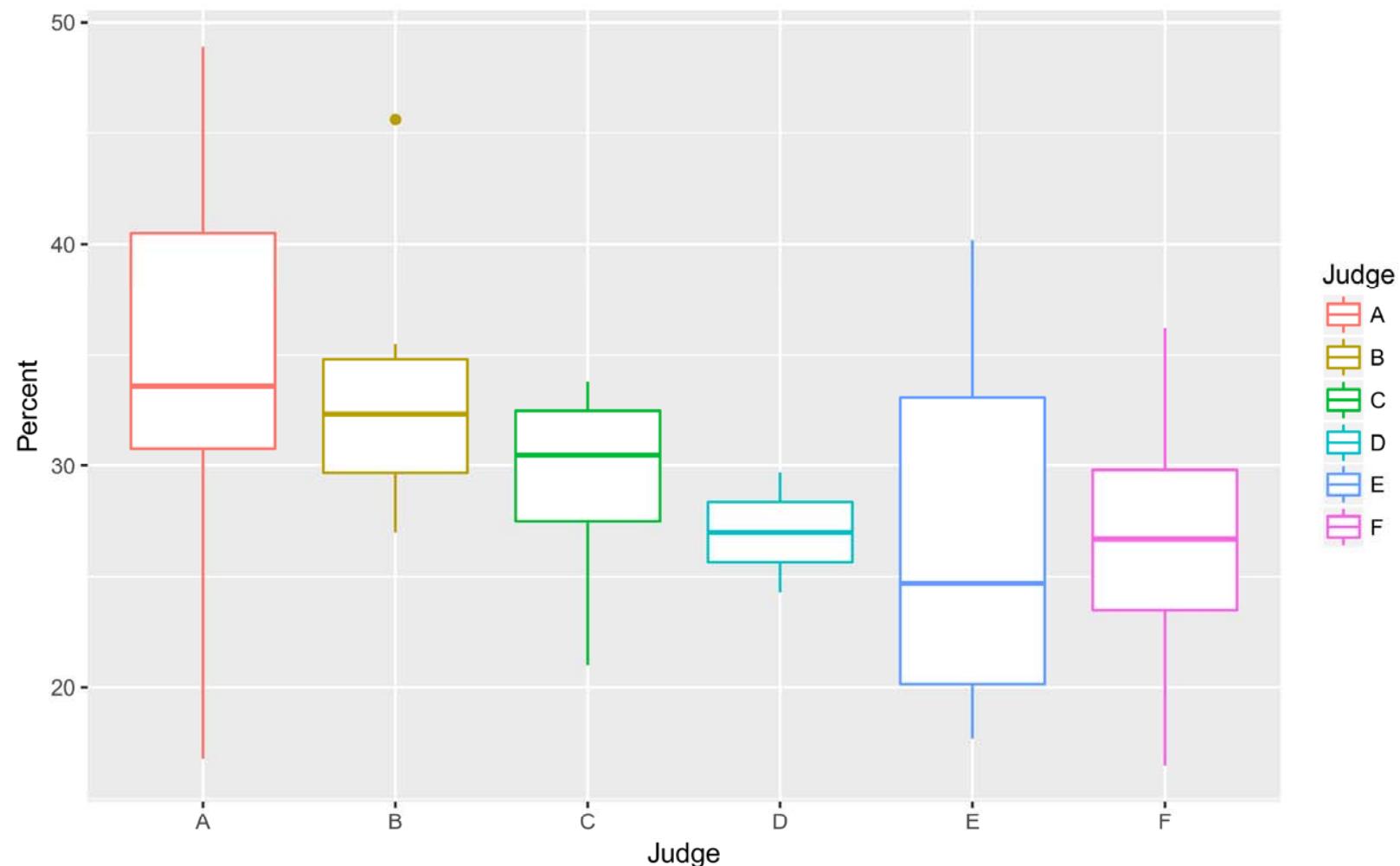
Case Study 1: Boxplot of Judges

```
# Get data subset of other judges
Others <- subset(jury, Judge != "Spock's")
boxplot(Percent~Judge, data=Others)
```



Case Study 1: Boxplot of Judges

```
#install.packages("ggplot2")
library(ggplot2)
ggplot(Others, aes(x=Judge,y=Percent, color=Judge))+geom_boxplot()
```



Case Study 1: Q2-Compare the 6 other judges

$$G = 6$$

```
summary(aov(Percent~Judge,data=Others))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
## Judge	5	326.5	65.29	1.218	0.324
## Residuals	31	1661.3	53.59		

MS Reg.

MSE

$$H_0: \mu_A = \mu_B = \dots = \mu_F$$

$$P\text{-value} = 0.324$$

- P-value is not small.
- We do not have evidence against the null hypoth.
- Data supports the idea that the variances of the other judges do not differ.

Case Study 1 Partial Summary

- (Q1) Data provides evidence that Spock's judge's venires underrepresent women.
 - ▶ Homoscedasticity satisfied
 - ▶ Normal errors hold
- (Q2) Data supports the hypothesis that the venires of the other six judges do not have similar percentages of women.
 - ▶ Where does the difference lie?
 - ▶ Are the model assumptions satisfied?

STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



January 16-18, 2018

One-way ANOVA

STA 303/1002: Week 2 Outline

- ▶ The General Linear Model
- ▶ One-way ANOVA
 - ▶ With $G=2$
 - ▶ With $G > 2$
- ▶ Case Study 1 continued
- ▶ Diagnostics- checking model assumptions
 - ▶ Normality of errors
 - ▶ Constant variance
 - ▶ Uncorrelated errors
- ▶ Multiple comparisons: Bonferroni and Tukey's

One-way ANOVA

The General Linear Model with Dummy Variables

One-way ANOVA

Simple Linear Model with 1 dummy variable:

$$\left\{ \begin{array}{l} H_0: \beta_1 = 0 = \mu_A - \mu_{A^c} \\ H_1: \mu_A = \mu_{A^c} \end{array} \right. \quad Y_i = \beta_0 + \beta_1 X_{i,A} + \epsilon_i \quad (G=2)$$
$$X_{i,A} = \begin{cases} 1 & \text{if } \text{ith obs. is from group A} \\ 0 & \text{o.w.} \end{cases}$$

Multiple Linear Model with G-1 dummy variables: $(G \geq 2)$

$$\left\{ \begin{array}{l} Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_{G-1} X_{i,G-1} + \epsilon_i \\ H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_{G-1} = 0 \\ \beta_1 = \mu_1 - \mu_g \quad \beta_2 = \mu_2 - \mu_g \quad \dots \quad \beta_{G-1} = \mu_{G-1} - \mu_g = 0 \end{array} \right.$$

One-way ANOVA
 $\Leftrightarrow \mu_1 = \mu_2 = \mu_3 = \dots = \mu_g$

One-way ANOVA Table

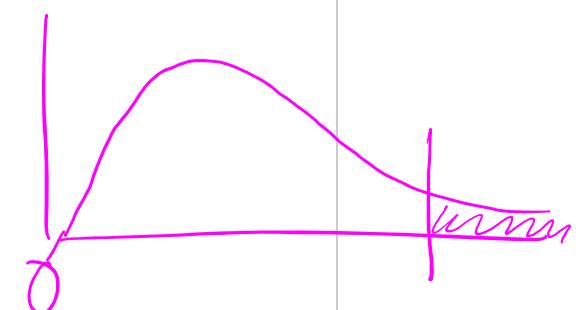
SOURCE	DF	SS	MS	F
Model	G-1	SSReg	MSReg=SSReg/G-1	MSReg/MSE
Error	N-G	RSS	MSE= RSS/N-G	
TOTAL	N-1	SST		

► Test statistic: $F = \frac{MS_{Reg}}{MSE}$

$$MSE = \hat{\sigma}^2 = s^2$$

► Distribution of test statistic: $\sim F_{G-1, N-G}$

► P-value: $P(F_{G-1, N-G} > F)$



General Linear Model vs One-way ANOVA

- ▶ General Linear Model:

- ▶ Response/Outcome, Y is continuous
- ▶ X 's are categorical and/or continuous
- ▶ Assumptions stated in terms of the errors, ie., $E_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$
- ▶ Assumptions are equivalent to One-way ANOVA
- ▶ In R: `lm()`

- ▶ One-way ANOVA

- ▶ Response/Outcome, Y is continuous
- ▶ One factor/categorical variable ($G \geq 2$)
- ▶ Assumptions are equivalent to General LM
- ▶ In R: `aov()`

Multiple Comparisons

- ▶ Post hoc procedure: further comparisons after significant result from overall One-way ANOVA
- ▶ 'Post hoc' means 'after this' in Latin
- ▶ Max of ${}^G C_2$ pairwise comparisons
- ▶ **Major issue:** There is an increased chance of making at least one Type I error when carrying out many tests.
- ▶ **Two common solutions:** based on controlling family Type I error rate
 - ▶ Bonferroni — controls α at pairwise level
 - ▶ Tukey's — controls α at family level

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$G = 3$$

$$k = 3 \quad C_2 = \binom{3}{2} = \frac{3!}{2! \cdot 1!} = 3$$

Pairs

1, 2
2, 3
1, 3

Multiple Comparisons

$$= P(\text{Type I Error})$$

Q: If $\alpha = 0.10$, what is the chance of committing 'at least 1' Type 1 Error...

- ▶ in 2 independent tests?

$$\begin{array}{cc} T_1 & T_2 \\ 1 - (1-\alpha)(1-\alpha) \end{array}$$

- ▶ in 10 independent tests?

- ▶ in k independent tests?

$$1 - (1-\alpha)^k$$

$$G=3, k=3$$

$$G=5, k=10$$

$$G=10, k=45$$

As $k \uparrow$, $(1-\alpha)^k \downarrow$, $1 - (1-\alpha)^k \uparrow$

Multiple Comparisons: Bonferroni's Method

$$(k=3)(\alpha=0.10) = 0.3.$$

- Based on the Bonferroni's inequality:

$$\underline{P(A \cup B)} \leq P(A) + P(B)$$

$$k\alpha$$

- Let A_i be the event that the i th test results in a Type I error.

$$\text{Then } P(UA_i) \leq \sum P(A_i)$$

$$k=45, \alpha=0.01$$

- Denote $k = {}^G C_2 = \binom{G}{2}$, total number of pairwise comparisons of G means.

$$k\alpha = 0.45$$

- Method: Conduct each of k pairwise tests at level $\underline{\alpha/k}$.

$$\boxed{\alpha^* = \alpha/k}$$

- Then the overall family Type I error rate of the k tests is at most α , i.e., the chance that at least 1 test results in a Type I error is at most α .

$$\alpha/k + \alpha/k + \dots + \alpha/k = \alpha.$$

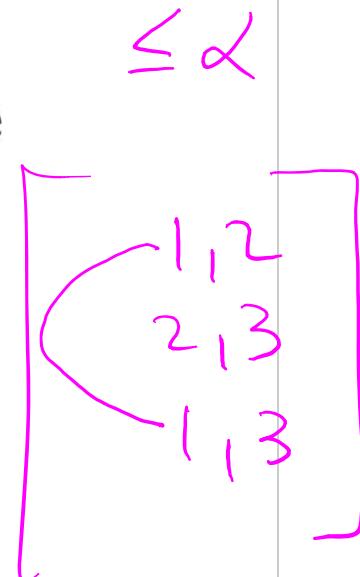
$$\leq \alpha.$$

Multiple Comparisons: Bonferroni's Method

- ▶ For CIs: If each CI has confidence level $(1 - \alpha/k)100\%$, then CI coverage rate is at most α

- ▶ Bonferroni CI: $|\bar{y}_i - \bar{y}_j| \pm t_{\frac{\alpha}{2k}}^* s_e(\bar{y}_i - \bar{y}_j)$

- ▶ Conservative: overall Type I error rate (chance of making at least one Type I error) is usually much less than α if tests are not mutually independent.
- ▶ Type II error inflation. Not as powerful.



1,2

3,4

5,6

Multiple Comparisons: Tukey's Approach

- ▶ Based on Tukey's Honestly Significant Difference (HSD)

- ▶ Requires Tukey's "Studentized Range Distribution" of $\max_{a,b \in \{1, \dots, G\}} \{\bar{y}_a - \bar{y}_b\}$

- ▶ Usually less conservative than Bonferroni's method, particularly if group sample sizes are similar.

*d v

- ▶ Precisely controls the overall Type I error rate at α ; simultaneous CI coverage rate is $(1 - \alpha)100\%$

Tukey's Approach: The Studentized Range distribution

- ▶ Consider n realizations- $\{x_1, \dots, x_n\}$ of a Normally distributed random variable, $X \sim N(\mu, \sigma^2)$. Determine the distribution of the largest and smallest value of $\{x_1, \dots, x_n\}$.
- ▶ Denote $\max\{X_1, \dots, X_n\} = X_{(n)}$ and $\min\{X_1, \dots, X_n\} = X_{(1)}$.
 $Range = X_{(n)} - X_{(1)}$.
- ▶ Based on n observations from X , the Studentized Range statistic is:
$$Q_{stat} = \frac{X_{(n)} - X_{(1)}}{s},$$
 s is the sample standard deviation
- ▶ Based on G group means, with n observations per group:

$$\bar{Q}_g = \frac{\sqrt{n} (\bar{y}_{(g)} - \bar{y}_{(1)})}{s_\nu},$$

where s_ν is the estimator of the pooled standard deviation, based on $\nu = N - G = G(n - 1)df$.

Significant Differences in 1-way ANOVA setting

- ▶ If there are G groups, then there is a maximum of $k = {}^G C_2$ pairwise differences.
- ▶ Controlling Overall/ Family/ Batch/ Experimentwise/ Simultaneaous Type I error rate versus Individual/ Comparisonwise/ Pairwise Type I error rate
- ▶ Finding a pairwise significant difference:
 - ▶ Compare method-wise Significant Difference, $c(\alpha)$ with $|\bar{y}_i - \bar{y}_j|$ OR
 - ▶ Determine whether confidence interval contains 0 OR
 - ▶ Compare P -value with α
- ▶ $s = \sqrt{MSE}$ with $df = \nu = \text{dfERROR}$

Sig diff $[H_0: M_i = M_j]$

One-way ANOVA

1. $c(\alpha)$ vs $|\bar{y}_i - \bar{y}_j|$
2. $P(2\text{-sided})$ vs α : $P < \alpha$
3. CI does not contain 0

Tukey's Honestly Significant Difference (HSD)

- ▶ Denote critical values from the Studentized Range distribution as $q(G, \nu, \alpha)$ or t^* .

► Family rate = α

► Pairwise rate $\leq \alpha$

$c(\alpha)$

Critical value from Tukey's distribution
One-way ANOVA

$$\text{Tukey's HSD} = \frac{q(G, \nu, \alpha)s}{\sqrt{n}}$$

γ means
if error
 $\sqrt{\text{MSE}}$

Similar to:

$$t^* \frac{s}{\sqrt{n}}$$

Bonferroni's significant differences

- ▶ Conduct each test at level α/k
 - ▶ Family rate $\leq \alpha$
 - ▶ Pairwise rate = α/k
 - ▶ Significant difference = $t_{\alpha/k, \nu} s \sqrt{\frac{2}{n}}$
- c(2) width of C-I .

Bartlett's Test for Homogeneity of Variances

- Extension of F-test for equality of 2 variances
- Hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_G^2$$

H_a : At least one σ_g^2 is different from the others

- Test statistic:

$$T = \frac{(N - G) \ln S_p^2 - \sum_{g=1}^G (n_g - 1) \ln S_g^2}{1 + \frac{1}{3(G-1)} \left(\sum_{g=1}^G \frac{1}{n_g - 1} - \frac{1}{N-G} \right)} \sim_{H_0} \chi_{G-1}^2$$

where $S_p^2 = \sum_g^G (n_g - 1) S_g^2 / (N - G)$ is the pooled variance

- In R: bartlett()
- A robust alternative test: Levene's, levene.test()

Model diagnostics: Any problems with model assumptions?

- ▶ **Homoscedasticity**: look at residuals in the diagnostic plots, use Bartlett's test, use rule of thumb
- ① ▶ **Normality**: use residual plots, or normal qq-plots
- ▶ Results: One slightly unusual observation but not influential value (large negative residual)

Model diagnostics: Constant variance?

②

- ▶ Constant variance: Rule of Thumb for variances

If $\frac{\text{largest } s_g}{\text{smallest } s_g} < 2$, assume variances are equal.

- ▶ For Spock's example, ignoring judge D since $n_D = 2$:

$$\frac{\text{largest } s_g}{\text{smallest } s_g} = \frac{11.9}{4.6} > 2$$

Hence, we may have a problem. Consider all inferences as only approximate.

③

- ▶ Uncorrelated errors: This is satisfied if venires are chosen independently.

Case Study I Conclusion

We have evidence that mean % women on venires is different between Spock's judge and all other judges except judge D ($n_D = 2$) and no evidence of difference among other judges.

Evidence of differences between
Spocks & A
" d B
" g C
" g E
" g F

$$H_0: M_S = M_D$$

Not
Rejected

Week 2 R functions

- ▶ One-way ANOVA: `aov()`
- ▶ Multiple Linear Regression Model: `summary(lm())`
- ▶ Barlett's Test of Equal Variance: `bartlett.test()`
- ▶ Bonferroni's: `pairwise.t.test()`, `confint()`
- ▶ Tukey's HSD: `TukeyHSD()`, `confint()`

One-way ANOVA

STA303/1004 - Week 2 R Markdown

January 16-18, 2018

Case Study 1: The Spock Conspiracy Trial Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case0502
library(Sleuth3)
#Juries data
jury = case0502
#attach(jury)
head(jury)
```

```
##   Percent    Judge
## 1     6.4 Spock's
## 2     8.7 Spock's
## 3    13.3 Spock's
## 4    13.6 Spock's
## 5    15.0 Spock's
## 6    15.2 Spock's
```

```
Percent=jury$Percent
Judge=jury$Judge
```

Compare variances of 6 other judges: Rule of thumb

```
others <- jury[Judge != "Spock's",]  
sss<-with(others, tapply(Percent,Judge,sd))  
sss  
  
##          A          B          C          D          E          F      Spock'  
## 11.941817  6.582224  4.592929  3.818377  9.010142  5.968878      N  
  
dim(sss)  
  
## [1] 7  
  
max(sss, na.rm=T)  
  
## [1] 11.94182  
  
min(sss, na.rm=T)  
  
## [1] 3.818377  
  
isTRUE((max(sss, na.rm=T)/min(sss, na.rm=T))>2)  
  
## [1] TRUE
```

Compare variances of 6 other judges: Bartlett's

```
bartlett.test(Percent~Judge, data=others)

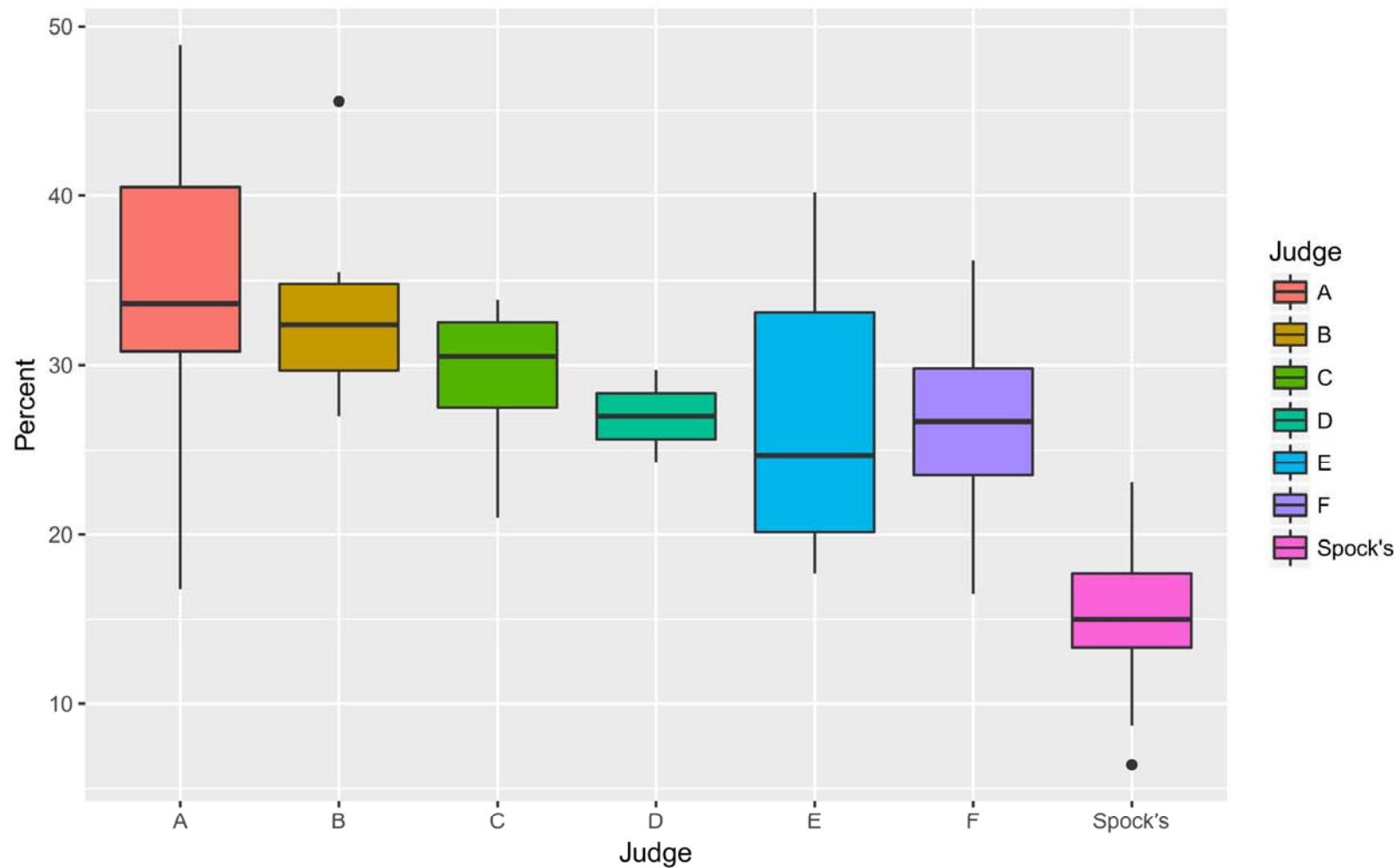
##
##  Bartlett test of homogeneity of variances
##
## data: Percent by Judge
## Bartlett's K-squared = 6.3125, df = 5, p-value = 0.277
```

Note: Group sizes are uneven and some are very small

Do not reject H_0
 H_0 assumed equal variances.

Compare all 7 judges

```
#boxplot(Percent~Judge)
library(ggplot2)
ggplot(jury, aes(x=Judge,y=Percent, fill=Judge)) +geom_boxplot()
```



Compare means of all 7 judges: One-way ANOVA

$$q=7$$

```
summary(aov(Percent~Judge))
```

```
##                Df Sum Sq Mean Sq F value    Pr(>F)    Small P-value
## Judge             6  1927   321.2   6.718 6.1e-05 ***  Sig. result.
## Residuals        39  1864    47.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compare means of all 7 judges: Gen Linear Model

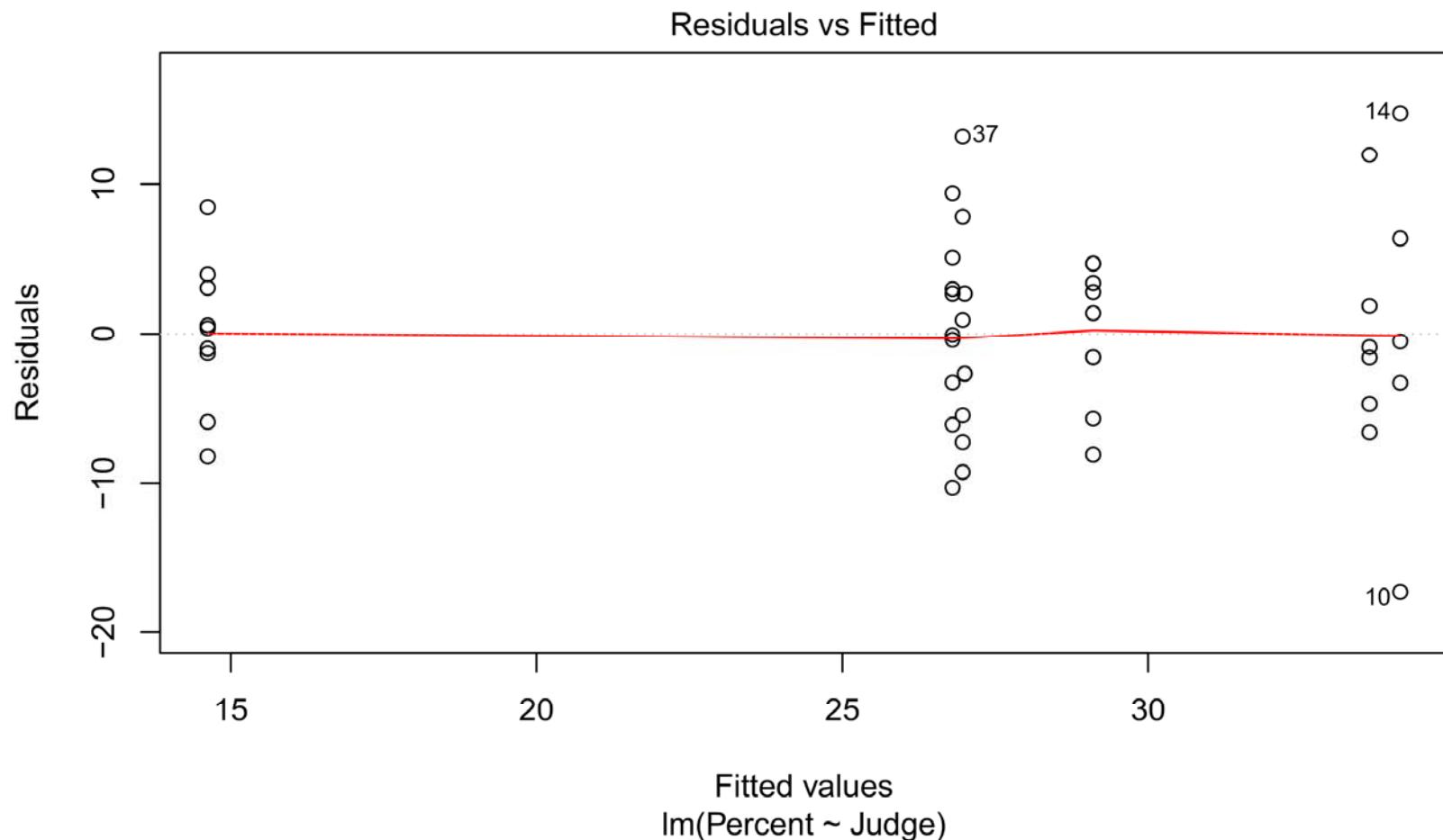
```
summary(lm(Percent ~ Judge))
```

```
##  
## Call:  
## lm(formula = Percent ~ Judge)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -17.320  -4.367  -0.250   3.319  14.780  
##  
## Coefficients: B  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  34.1200   3.0921  11.034 1.47e-13 ***  
## JudgeB     -0.5033   4.1868  -0.120  0.9049  
## JudgeC     -5.0200   3.8566  -1.302  0.2007  
## JudgeD    -17.1200   5.7848  -1.231  0.2258  
## JudgeE    -17.1533   4.1868  -1.709  0.0955 .  
## JudgeF    -17.3200   3.8566  -1.898  0.0651 .  
## JudgeSpock's -19.4978   3.8566  -5.056 1.05e-05 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 6.914 on 39 degrees of freedom
```

Pr(>|t|) as table
in ANOVA table
in R previous page
Same in ANOVA table
in R previous page

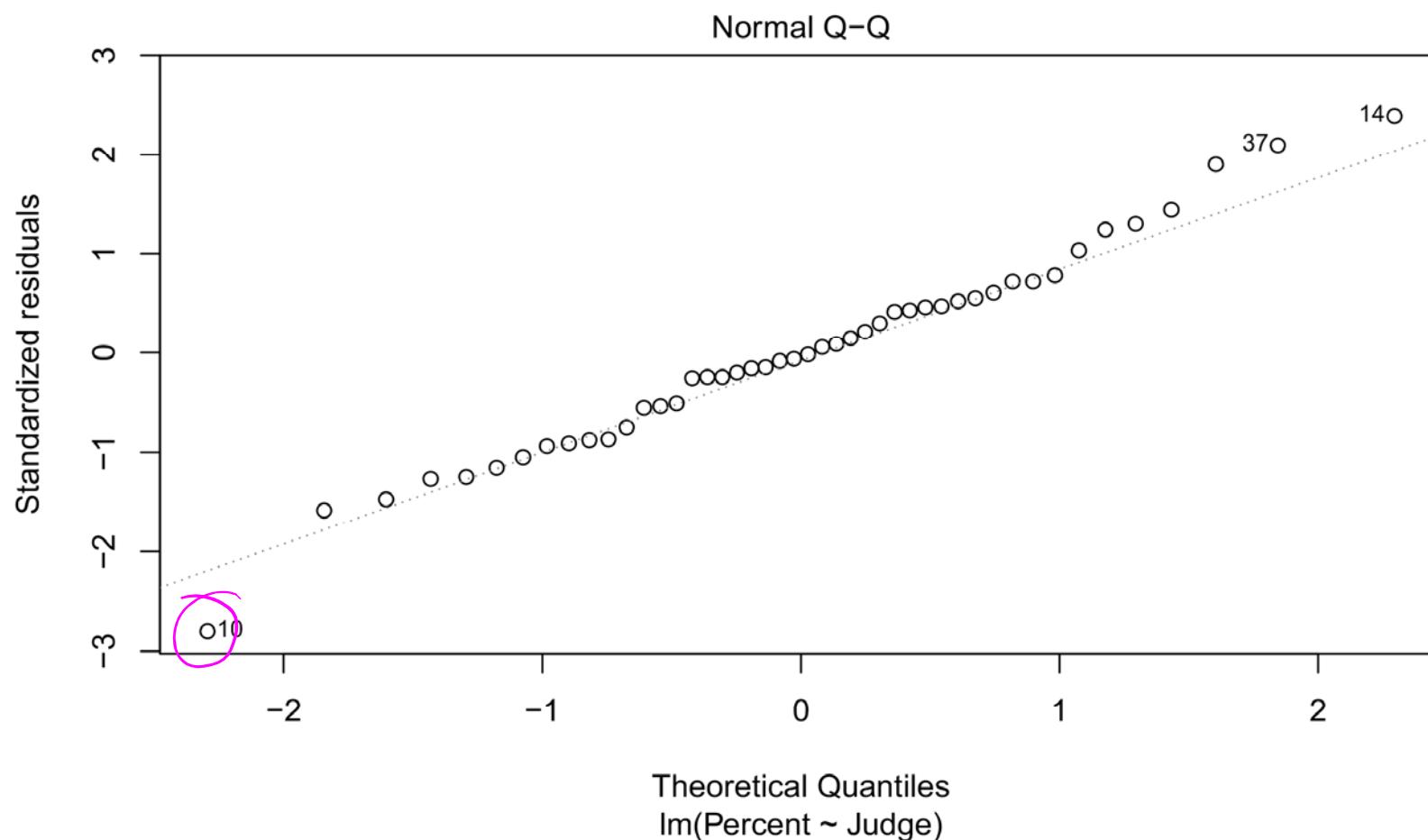
Check Normality: Linear Model

```
plot(lm(Percent~Judge), which=1)
```



Check Normality: Linear Model

```
plot(lm(Percent~Judge), which=2)
```



Compare variances of all 7 judges: RoT

```
ssa=with(jury, tapply(Percent, Judge, sd))
ssa

##           A            B            C            D            E            F        Spock'
## 11.941817  6.582224  4.592929  3.818377  9.010142  5.968878  5.03879

isTRUE((max(ssa, na.rm=T)/min(ssa, na.rm=T))>2)
```

[1] TRUE

Compare variances of all 7 judges: Bartlett's

```
bartlett.test(Percent~Judge, data=jury)

##
##  Bartlett test of homogeneity of variances
##
## data: Percent by Judge
## Bartlett's K-squared = 7.7582, df = 6, p-value = 0.2564
```

Case Study 1: Bonferroni's

```
Judge=relevel(Judge, ref="Spock's")
pairwise.t.test(Percent, Judge, p.adj="bonf")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: Percent and Judge
##
## Spock's A B C D E
## A 0.00022 - - - -
## B 0.00013 1.00000 - - -
## C 0.00150 1.00000 1.00000 - - -
## D 0.57777 1.00000 1.00000 1.00000 - -
## E 0.03408 1.00000 1.00000 1.00000 1.00000 -
## F 0.01254 1.00000 1.00000 1.00000 1.00000 1.00000
##
## P value adjustment method: bonferroni
```

$$\mu_A - \mu_S$$

$$\mu_B - \mu_S$$

:

large P-value (0.5777)
for the test of $H_0: \mu_D - \mu_{\text{Spock}} = 0$
that μ_D is similar
to μ_{Spock}
⇒ Evidence

Case Study 1: Bonferroni's CIs

```
lmod=lm(Percent~Judge)
nlevels(jury$Judge)

## [1] 7

confint(lmod, level=1-0.05/nlevels(jury$Judge))

##               0.357 % 99.643 %
## (Intercept) 8.078085 21.16636
## JudgeA      8.547341 30.44821
## JudgeB      8.647255 29.34163
## JudgeC      5.222970 23.73259
## JudgeD      -2.969585 27.72514
## JudgeE      1.997255 22.69163
## JudgeF      2.922970 21.43259
```

Includes 0

Case Study 1: Bonferroni's CIs

```
summary(lmod)
```

```
##  
## Call:  
## lm(formula = Percent ~ Judge)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -17.320  -4.367  -0.250   3.319  14.780  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 14.622     2.305   6.344 1.72e-07 ***  
## JudgeA     19.498     3.857   5.056 1.05e-05 ***  
## JudgeB     18.994     3.644   5.212 6.39e-06 ***  
## JudgeC     14.478     3.259   4.442 7.15e-05 ***  
## JudgeD     12.378     5.405   2.290 0.027513 *  
## JudgeE     12.344     3.644   3.388 0.001623 **  
## JudgeF     12.178     3.259   3.736 0.000597 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 6.914 on 39 degrees of freedom
```

Case Study 1: Tukey's CIs

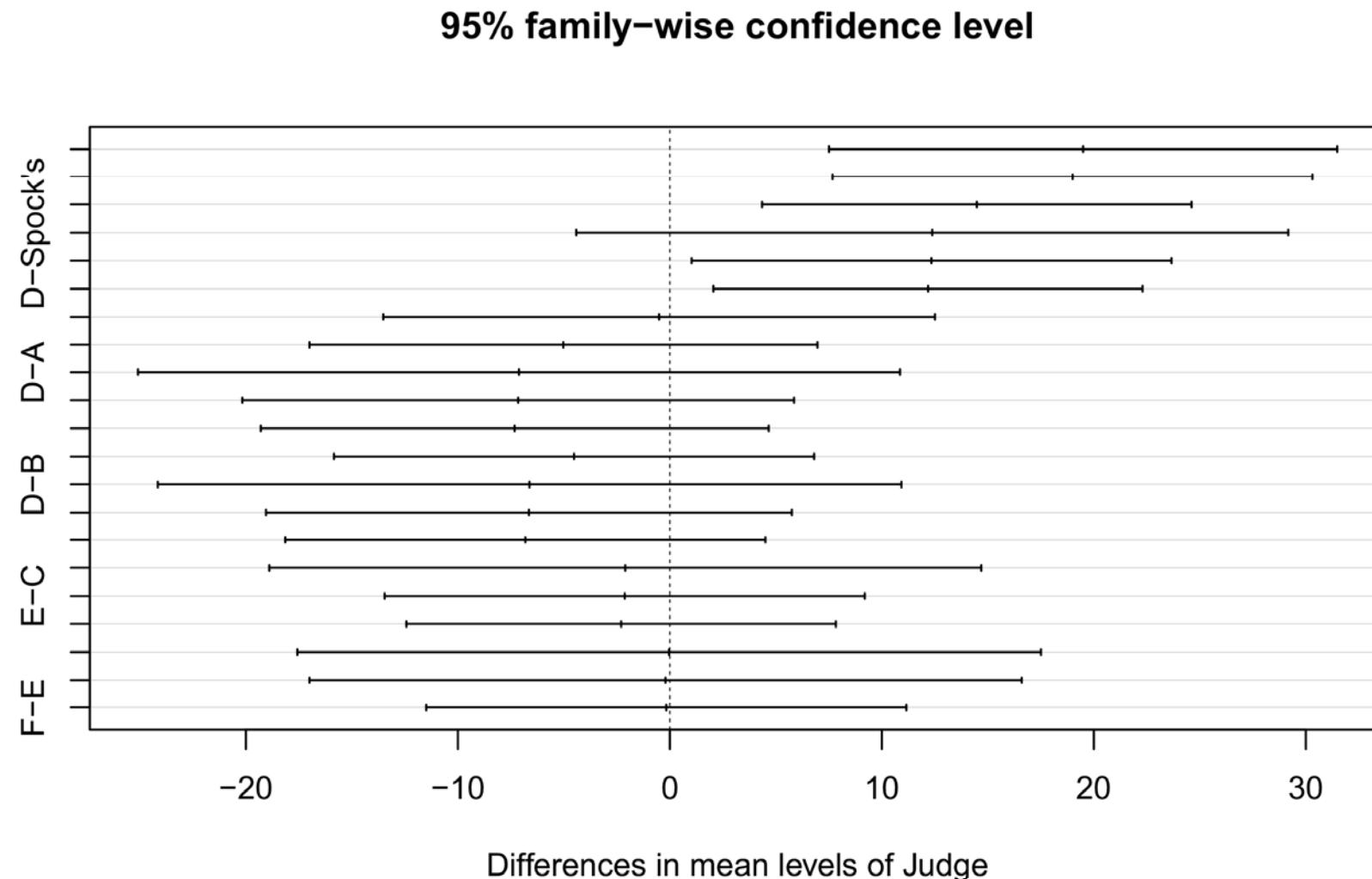
```
amod=aov(Percent~Judge)
TukeyHSD(amod, "Judge")
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Percent ~ Judge)
##
## $Judge
##          diff      (lwr       upr )   p adj
## A-Spock's 19.4977778 7.514686 31.480870 0.0001992
## B-Spock's 18.9944444 7.671487 30.317402 0.0001224
## C-Spock's 14.4777778 4.350216 24.605339 0.0012936
## D-Spock's 12.3777778 -4.416883 29.172438 0.2744263
## E-Spock's 12.3444444 1.021487 23.667402 0.0248789
## F-Spock's 12.1777778 2.050216 22.305339 0.0098340
## B-A      -0.5033333 -13.512422 12.505755 0.9999997
## C-A      -5.0200000 -17.003092  6.963092 0.8470097
## D-A      -7.1200000 -25.094638 10.854638 0.8777485
## E-A      -7.1533333 -20.162422  5.855755 0.6146238
## F-A      -7.3200000 -19.303092  4.663092 0.4936379
## C-B      -4.5166667 -15.839625  6.806291 0.8742030
## D-B      -6.6166667 -24.158118 10.924784 0.9003280
```

k=21

Case Study 1: Tukey's CI's

```
plot(TukeyHSD(amod, "Judge"))
```



A, S
B, S
C, S
E, S
F, S

STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



January 23-25, 2018

Two-way ANOVA

STA 303/1002: Week 3 Intro

- ▶ Review: General Linear Model (GLM)
 - ▶ Response, Y is continuous
 - ▶ Categorical or continuous predictors, X
 - ▶ Y is linear in β 's
 - ▶ Assumptions: $\epsilon \sim N(0, \sigma^2 I)$
- ▶ Review: One-way ANOVA
 - ▶ Special case of a GLM
 - ▶ One-way classification/ One factor with $G \geq 2$ levels
- ▶ What if we have more than one factor?
 - ▶ Main and Interaction effect of factors on Y ?
 - ▶ Assumptions?
 - ▶ Visualizations?
 - ▶ Analyses?

Two-way ANOVA

Two-way Classification or Two-way Analysis of Variance

General LM

- ▶ Another special case of a GLM
- ▶ Extension of One-way ANOVA
- ▶ Two factors, each with at least 2 levels ($G_1 \geq 2, G_2 \geq 2$)
- ▶ Uses a maximum of $(G_1 - 1) + (G_2 - 1) + (G_1 - 1)(G_2 - 1)$
indicator variables

Terminology from Design of Experiments:

- ▶ **Factor**- a categorical predictor variable, eg. *Treatment*
- ▶ Factors are composed of different class levels, eg. various types of treatments

Two-way ANOVA

Two types of factors

- ▶ FIXED effect: data has been gathered from all the levels of the factor that are of interest
- ▶ Random effect: interest is in all possible levels of the factor, but only a random sample of levels is included in the data
- ▶ Egs.: Suppose measurements are taken on the yield of a machine operated by each of several operators. We want to compare the mean yields under different operators.
 - ▶ Factor: operator
 - ▶ Fixed effect: Interest is only in those particular operators (may be all the operators at the plant)
 - ▶ Random effect: Operators are a random sample from larger population of all operators.

Case Study II-The Pygmalion Effect

- ▶ *Pygmalion effect*- high expectations of a supervisor or teacher translate to improved performance by subordinates or students
- ▶ Data:

Company	<u>Treatments</u>		
	Pygmalion	Control	
1 (3)	80.0	63.2	69.2
2	83.9	63.1	81.5
3 (2)	68.2	76.2	
4	76.5	59.5	73.5
5	87.8	73.9	78.5
6	89.8	78.9	84.7
7	76.1	60.6	69.6
8	71.5	67.8	73.2
9	69.5	72.3	73.9
10	83.7	63.7	77.7

Two-way ANOVA

29 obs.

Case Study II-The Pygmalion Effect

- ▶ Setup:

- ▶ A randomized experiment to test Pygmalion effect
- ▶ Used 10 companies in an army training camp
- ▶ Most companies have 3 platoons; each platoon trains together under 1 leader (1 leader per platoon).
- ▶ Within each company, 1 platoon leader was told that he has an exceptionally good group- this is the pygmalion platoon; the other 2 are control platoons.
- ▶ Each pygmalion platoon was randomly chosen.

- ▶ Experimental units: platoons

1, ..., 29

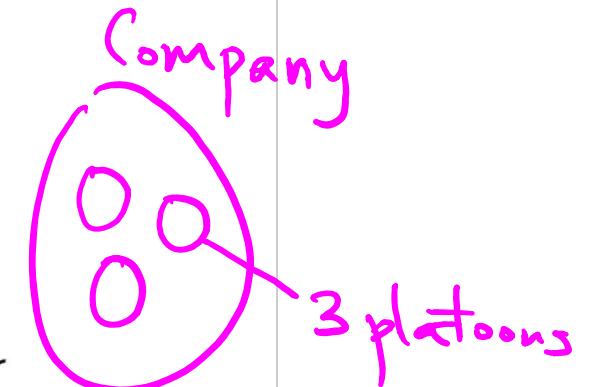
- ▶ Unbalanced design: one company had only two platoons



- ▶ Response: score on a basic weapons test per platoon

- ▶ Factors:

- (1) *Company*- 10 levels (company 1, ..., company 10)
- (2) *Treatment*- 2 levels (pygmalion, control)



Case Study II Objective

- ▶ **Aim:** Investigate the interaction between *Company* and *Treatment*
- ▶ **Method:** Fit a Two-way ANOVA (a General LM)

Two-way ANOVA

Case Study II Variables

- ▶ Response: Y_i - score for i th platoon, $i = 1, \dots, 29$
- ▶ Explanatory variables: $9 + 1 + 9$ Indicator variables
 - ▶ 9 for Company ($\mathbb{1}_{COMP_1,i}, \dots, \mathbb{1}_{COMP_9,i}$)
 - ▶ 1 for Treatment ($\mathbb{1}_{PYG,i}$)
 - ▶ 9 for interaction terms
 $(\mathbb{1}_{PYG,i} \times \mathbb{1}_{COMP_1,i}, \dots, \mathbb{1}_{PYG,i} \times \mathbb{1}_{COMP_9,i})$

where

$$\mathbb{1}_{PYG,i} = \begin{cases} 1 & \text{if } i\text{th platoon is "pygmalion"} \\ 0 & \text{if } i\text{th platoon is "control"} \end{cases}$$

$$\mathbb{1}_{COMP_1,i} = \begin{cases} 1 & \text{if } i\text{th platoon is from "company 1"} \\ 0 & \text{if } i\text{th platoon is NOT from "company 1"} \end{cases}$$

Case Study II Linear Model

Full Model:

$$Y_i = \beta_0 + \beta_1 \mathbb{1}_{PYG,i} + \beta_2 \mathbb{1}_{COMP_1,i}$$

$$+ \beta_3 \mathbb{1}_{COMP_2,i}$$

$$+ \dots$$

$$+ \beta_{10} \mathbb{1}_{COMP_9,i}$$

1st platoon, PYG, Comp 1

$$\begin{aligned} &+ \beta_{11} \mathbb{1}_{PYG,i} \times \mathbb{1}_{COMP_1,i} \\ &+ \beta_{12} \mathbb{1}_{PYG,i} \times \mathbb{1}_{COMP_2,i} \\ &+ \dots \\ &+ \beta_{19} \mathbb{1}_{PYG,i} \times \mathbb{1}_{COMP_9,i} \\ &+ \epsilon_i \end{aligned}$$

$$E[y_i | (treat_i, company_i)] = \beta_0 + \beta_1 + \beta_2 + \beta_{11}$$

Two-way ANOVA

Case Study II: Expected Response | (Company*Treatment)

(Pyg - Control)

Company	$\text{Pygmalion}(\mathbb{1}_{PYG,i} = 1)$	$\text{Control}(\mathbb{1}_{PYG,i} = 0)$	Treatment effect
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{11}$	$\beta_0 + \beta_2$	$\beta_1 + \beta_{11}$
2	$\beta_0 + \beta_1 + \beta_3 + \beta_{12}$	$\beta_0 + \beta_3$	$\beta_1 + \beta_{12}$
3			
4			
5			
6			
7			
8			
9	$\beta_0 + \beta_1 + \beta_{10} + \beta_{19}$	$\beta_0 + \beta_{10}$	$\beta_1 + \beta_{19}$
10	$\beta_0 + \beta_1$	β_0	β_1

Question 1: Does mean treatment effect differ with Company?

Null Hypothesis, H_0 : $\beta_{11} = \beta_{12} = \beta_{13} = \dots = \beta_{19} = 0$

Alternative Hypothesis, H_a :

at least 1 β is not 0

Two-way ANOVA

Overall versus Partial F-tests in Two-way ANOVA

Full

- Overall test: $H_0 : \beta_1 = \beta_2 = \dots = \beta_{df\text{MODEL}} = 0$

Reduced

- Partial test: $H_0 : \text{a subset of } \{\beta_1, \beta_2, \dots, \beta_{df\text{MODEL}}\} = 0$
- Approach: Fit full model (with all explanatory variables) and reduced (without variables whose coefficients you are testing) model

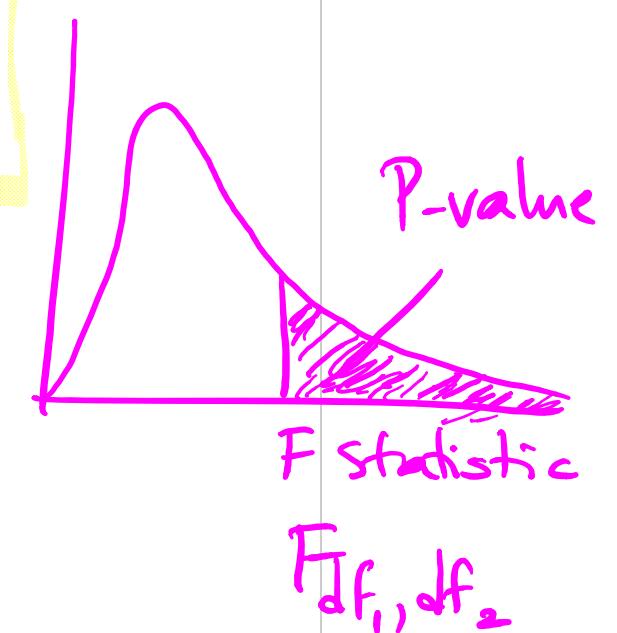
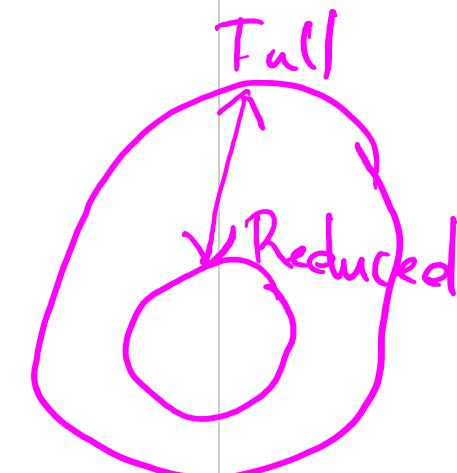
- Test statistic:

$$F = \frac{(SSReg_{full} - SSReg_{reduced}) / (\# \text{ of } \beta's \text{ - being - tested})}{MSE_{full}}$$

$$= \frac{(RSS_{reduced} - RSS_{full}) / (\# \text{ of } \beta's \text{ - being - tested})}{MSE_{full}}$$

- If H_0 is true, F is an observation from F distribution with $df = (\# \text{ of } \beta's \text{ being tested}, df\text{ERROR of full model})$

Two-way ANOVA



Case Study II: Testing interaction

- ▶ FULL:

full=lm(score~company*treat)

19

- ▶ Reduced:

reduced=lm(score~company+treat)

10

$$19 - 10 = 9$$

$\beta_{11}, \beta_{12}, \dots, \beta_{19}$

- ▶ Partial F-test (Refer to R output)

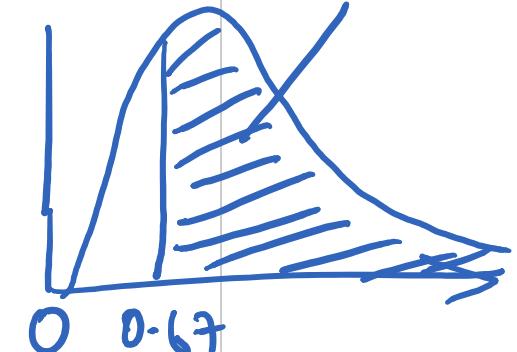
- ▶ Test statistic:

$$F = \frac{(1321.32 - 1009.86)/9}{51.89} = \frac{(778.5 - 467.04)/9}{51.89} = \frac{311.46/9}{51.89} = 0.67$$

P-value

- ▶ Under H_0 , F statistic $\sim F$ distribution with $df = (9, 9)$.
- ▶ The resulting p -value is large ($p = 0.7221$), implying that the data are consistent with zero coefficient for the interaction term.
- ▶ No evidence that treatment effect differs with Company.

⇒ Fit Additive model



$F_{9,9}$

Case Study II: Interaction model summary

```
Call:  
lm(formula = Score ~ company * treat)  
  
Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 66.200 5.094 12.996 3.89e-07 ***  
companyC10 4.500 7.204 0.625 0.5477  
companyC2 6.100 7.204 0.847 0.4191  
companyC3 10.000 8.823 1.133 0.2863  
companyC4 0.300 7.204 0.042 0.9677  
companyC5 10.000 7.204 1.388 0.1985  
companyC6 15.600 7.204 2.166 0.0585 .  
companyC7 -1.100 7.204 -0.153 0.8820  
companyC8 4.300 7.204 0.597 0.5653  
companyC9 6.900 7.204 0.958 0.3632  
treatPygmalion 13.800 8.823 1.564 0.1522  
companyC10:treatPygmalion -0.800 12.477 -0.064 0.9503  
companyC2:treatPygmalion -2.200 12.477 -0.176 0.8639  
companyC3:treatPygmalion -21.800 13.477 -1.618 0.1402  
companyC4:treatPygmalion -3.800 12.477 -0.305 0.7676  
companyC5:treatPygmalion -2.200 12.477 -0.176 0.8639  
companyC6:treatPygmalion -5.800 12.477 -0.465 0.6531  
companyC7:treatPygmalion -2.800 12.477 -0.224 0.8275  
companyC8:treatPygmalion -12.800 12.477 -1.026 0.3317  
companyC9:treatPygmalion -17.400 12.477 -1.395 0.1966  
  
Residual standard error: 7.204 on 9 degrees of freedom  
Multiple R-squared: 0.7388, Adjusted R-squared: 0.1875  
F-statistic: 1.34 on 19 and 9 DF, p-value: 0.3358
```

Two-way ANOVA

Case Study II: Additive Model

Additive (a reduced) Model:

$$Y_i = \beta_0 + \beta_1 \mathbb{1}_{PYG,i} + \beta_2 \mathbb{1}_{COMP_1,i} + \beta_3 \mathbb{1}_{COMP_2,i} + \dots + \beta_{10} \mathbb{1}_{COMP_9,i} + \epsilon_i$$

Treat

Company

Two-way ANOVA

Case Study II: Additive Model Expected Response

Company	<i>Treatment</i>		Treatment effect
	Pygmalion($\mathbb{1}_{PYG,i} = 1$)	Control($\mathbb{1}_{PYG,i} = 0$)	
1	$\beta_0 + \beta_1 + \beta_2$	$\beta_0 + \beta_2$	β_1
2	$\beta_0 + \beta_1 + \beta_3$	$\beta_0 + \beta_3$	β_1
...
8	$\beta_0 + \beta_1 + \beta_9$	$\beta_0 + \beta_9$	β_1
9	$\beta_0 + \beta_1 + \beta_{10}$	$\beta_0 + \beta_{10}$	β_1
10	$\beta_0 + \beta_1$	β_0	β_1

Test 1: Is there a difference in mean score between pygmalion and control group? $H_0: \beta_1 = 0 = \mu_{PYG} - \mu_{control}$

Test 2: Are there differences between companies?

$$H_0: \beta_2 = \beta_3 = \dots = \beta_{10} = 0$$

Case Study II: Additive model summary

Call:

lm(formula = Score ~ company + treat)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	68.39316	3.89308	17.568	8.92e-13 ***
companyC10	4.23333	5.36968	0.788	0.4407
companyC2	5.36667	5.36968	0.999	0.3308
companyC3	0.19658	6.01886	0.033	0.9743
companyC4	-0.96667	5.36968	-0.180	0.8591
companyC5	9.26667	5.36968	1.726	0.1015
companyC6	13.66667	5.36968	2.545	0.0203 *
companyC7	-2.03333	5.36968	-0.379	0.7094
companyC8	0.03333	5.36968	0.006	0.9951
companyC9	1.10000	5.36968	0.205	0.8400
treatPygmalion	7.22051	2.57951	2.799	0.0119 *

Residual standard error: 6.576 on 18 degrees of freedom

Multiple R-squared: 0.5647, Adjusted R-squared: 0.3228

F-statistic: 2.335 on 10 and 18 DF, p-value: 0.0564

Two-way ANOVA

Case Study II: Additive model summary

Call:

```
lm(formula = Score ~ treat + company)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	68.39316	3.89308	17.568	8.92e-13 ***
treatPygmalion	7.22051	2.57951	2.799	0.0119 *
companyC10	4.23333	5.36968	0.788	0.4407
companyC2	5.36667	5.36968	0.999	0.3308
companyC3	0.19658	6.01886	0.033	0.9743
companyC4	-0.96667	5.36968	-0.180	0.8591
companyC5	9.26667	5.36968	1.726	0.1015
companyC6	13.66667	5.36968	2.545	0.0203 *
companyC7	-2.03333	5.36968	-0.379	0.7094
companyC8	0.03333	5.36968	0.006	0.9951
companyC9	1.10000	5.36968	0.205	0.8400

Residual standard error: 6.576 on 18 degrees of freedom

Multiple R-squared: 0.5647, Adjusted R-squared: 0.3228

F-statistic: 2.335 on 10 and 18 DF, p-value: 0.0564

Two-way ANOVA

Case Study II: Additive Model-Testing main effects

	Test 1	Test 2
Null	$H_0 : \beta_1 = 0$	$H_0 : \beta_2 = \beta_3 = \dots = \beta_{10} = 0$
Alt	$H_a : \beta_1 \neq 0$	$H_a : \text{at least one } \beta \neq 0$
F statistic	7.84	1.75
F -dist df	(1,18)	(9,18)
p -value	0.0119	0.1484
Conc.	Evidence of a difference in mean score between pygmalion and control platoons (over and above difference btw companies)	No evidence of difference between companies.

- ▶ On average, pygmalion platoons (mean=78.7) scored higher than control platoons (mean=71.6). ↗

Case Study II: Model Checking

- ▶ Look at diagnostic panel of plots
 - ▶ No outliers
 - ▶ Normality ok
 - ▶ Perhaps decreasing variance
- ▶ Independent observations: by assuming that platoons were chosen at random and were not interacting

STA303/1004 - Week 3 R Markdown

January 23-25, 2018

Case Study 2: The Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case1302
library(Sleuth3)
#Pygmalion data
pyg = case1302
attach(pyg)
head(pyg)
```

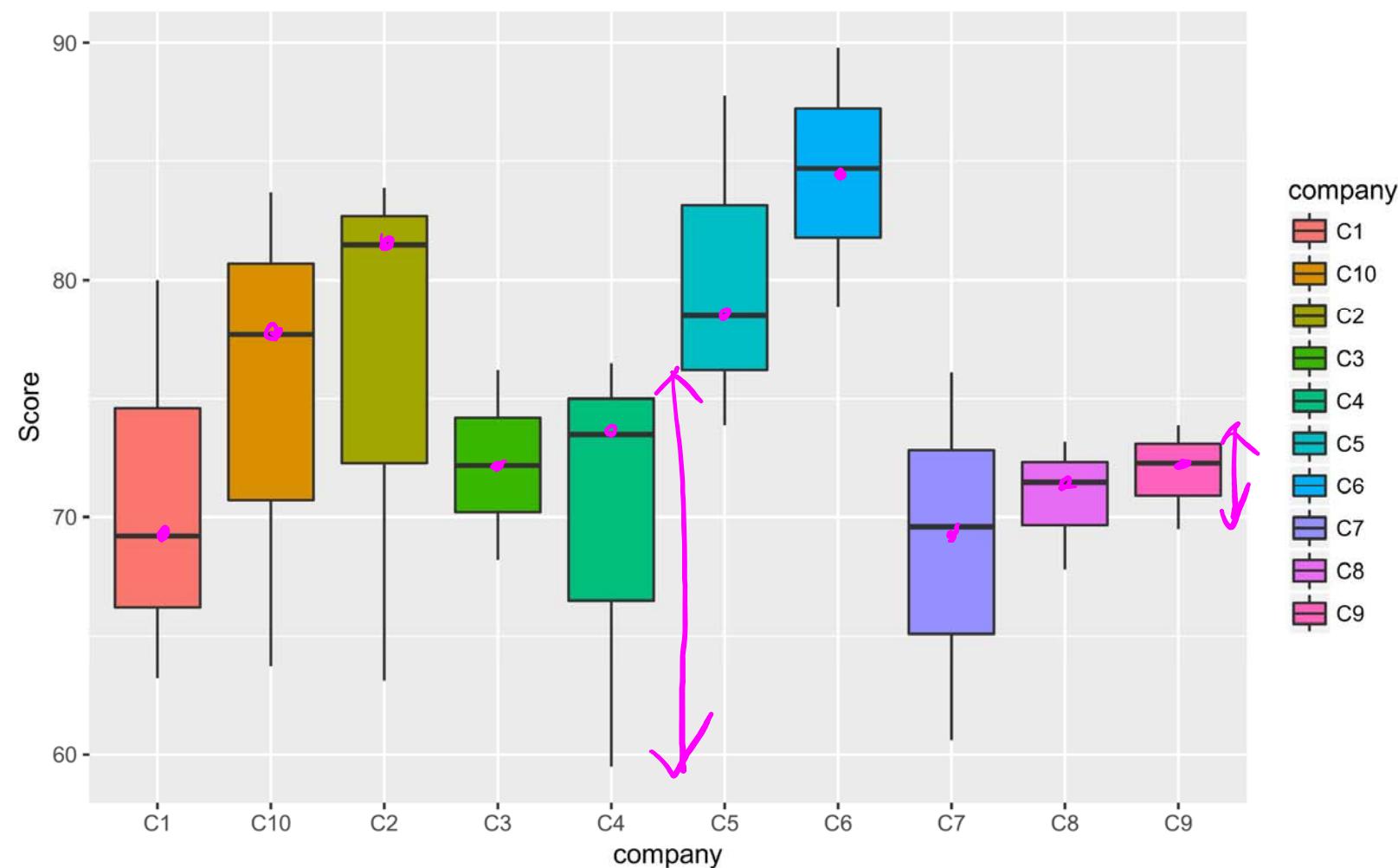
```
##   Company    Treat Score
## 1      C1 Pygmalion  80.0
## 2      C1   Control  63.2
## 3      C1   Control  69.2
## 4      C2 Pygmalion  83.9
## 5      C2   Control  63.1
## 6      C2   Control  81.5
```

29 ↓

```
company=as.factor(Company)
treat=as.factor(Treat)
```

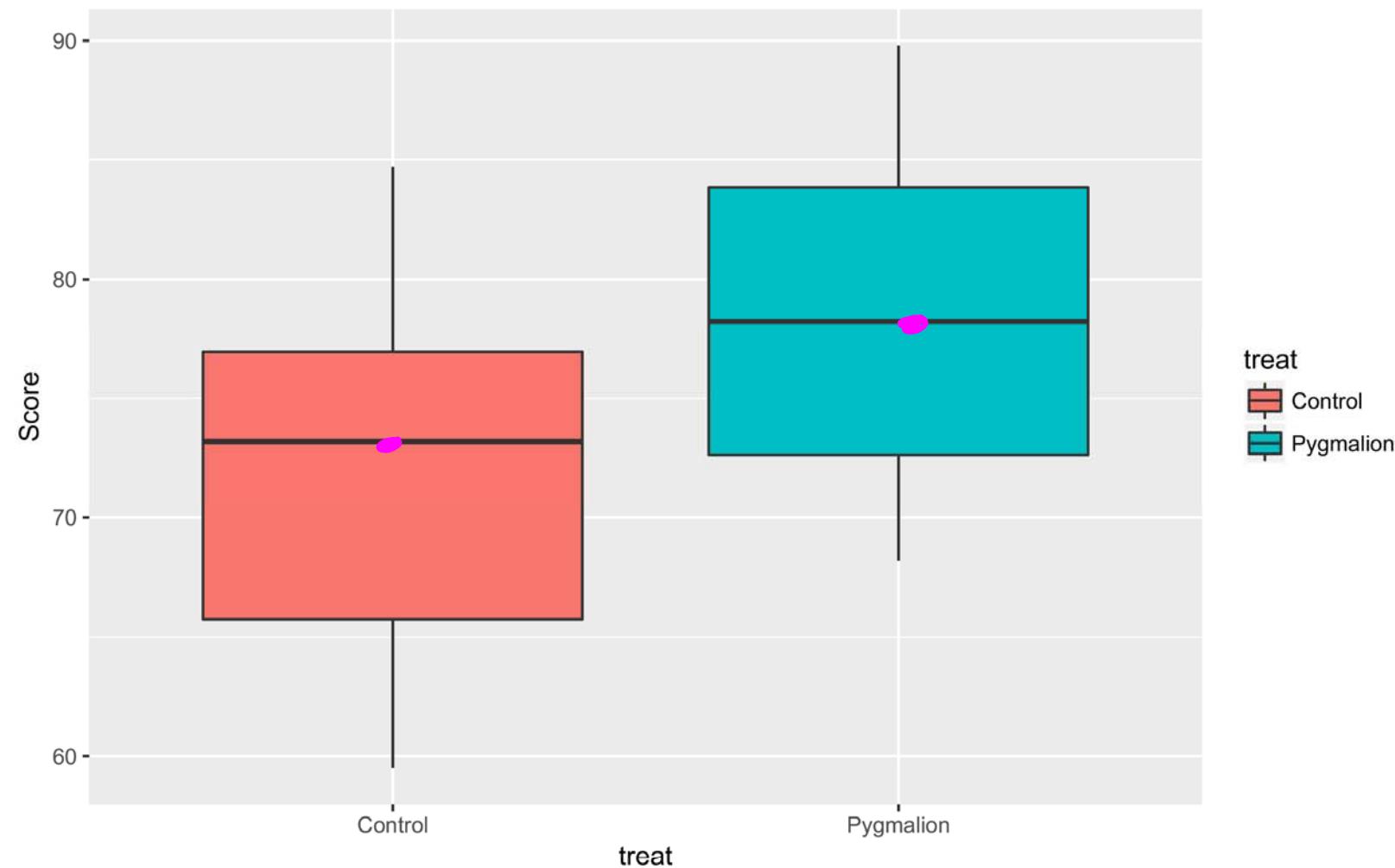
Case Study 2: Visualizing the data

```
#install.packages("ggplot2")
library(ggplot2)
pc<-ggplot(pyg, aes(x=company,y=Score, fill=company))+geom_boxplot()
pc
```



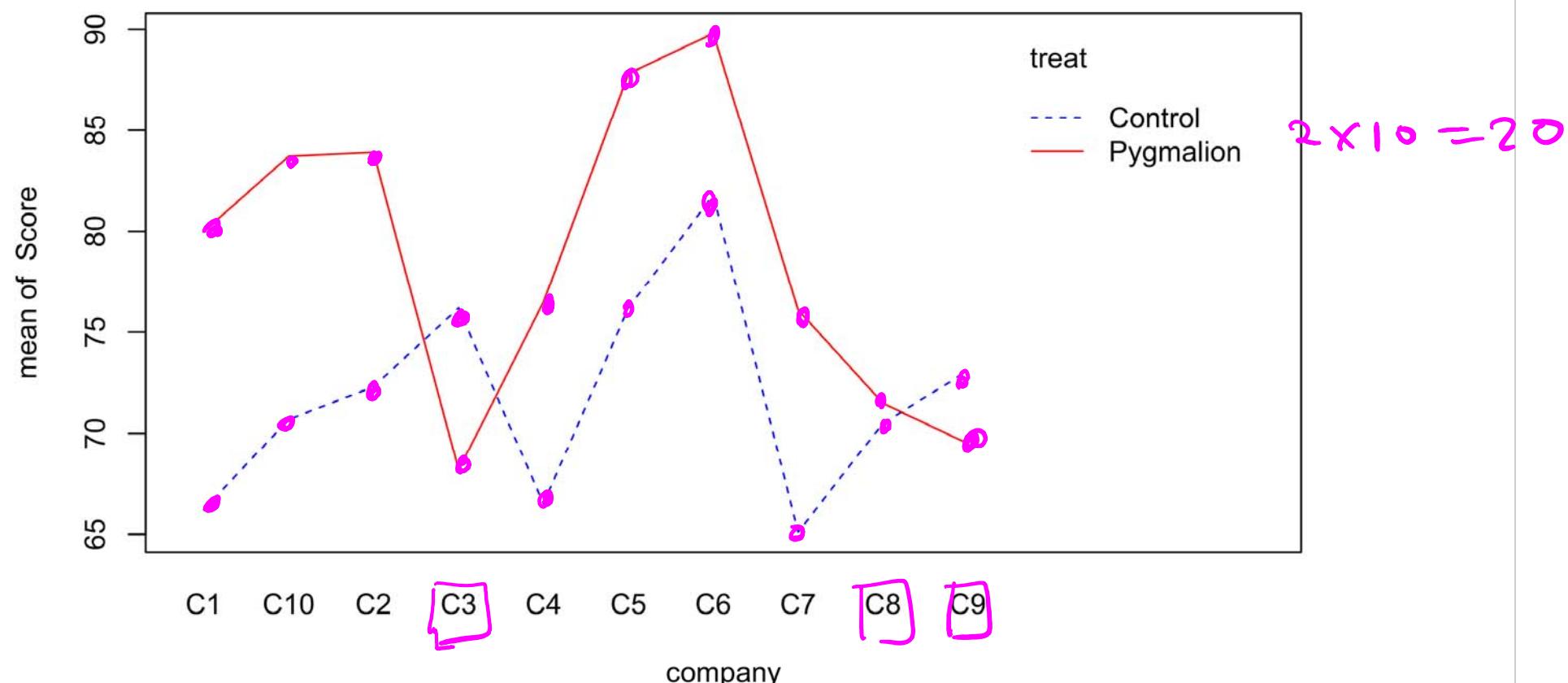
Case Study 2: Visualizing the data

```
ptr<-ggplot(pyg, aes(x=treat,y=Score, fill=treat))+geom_boxplot()  
ptr
```



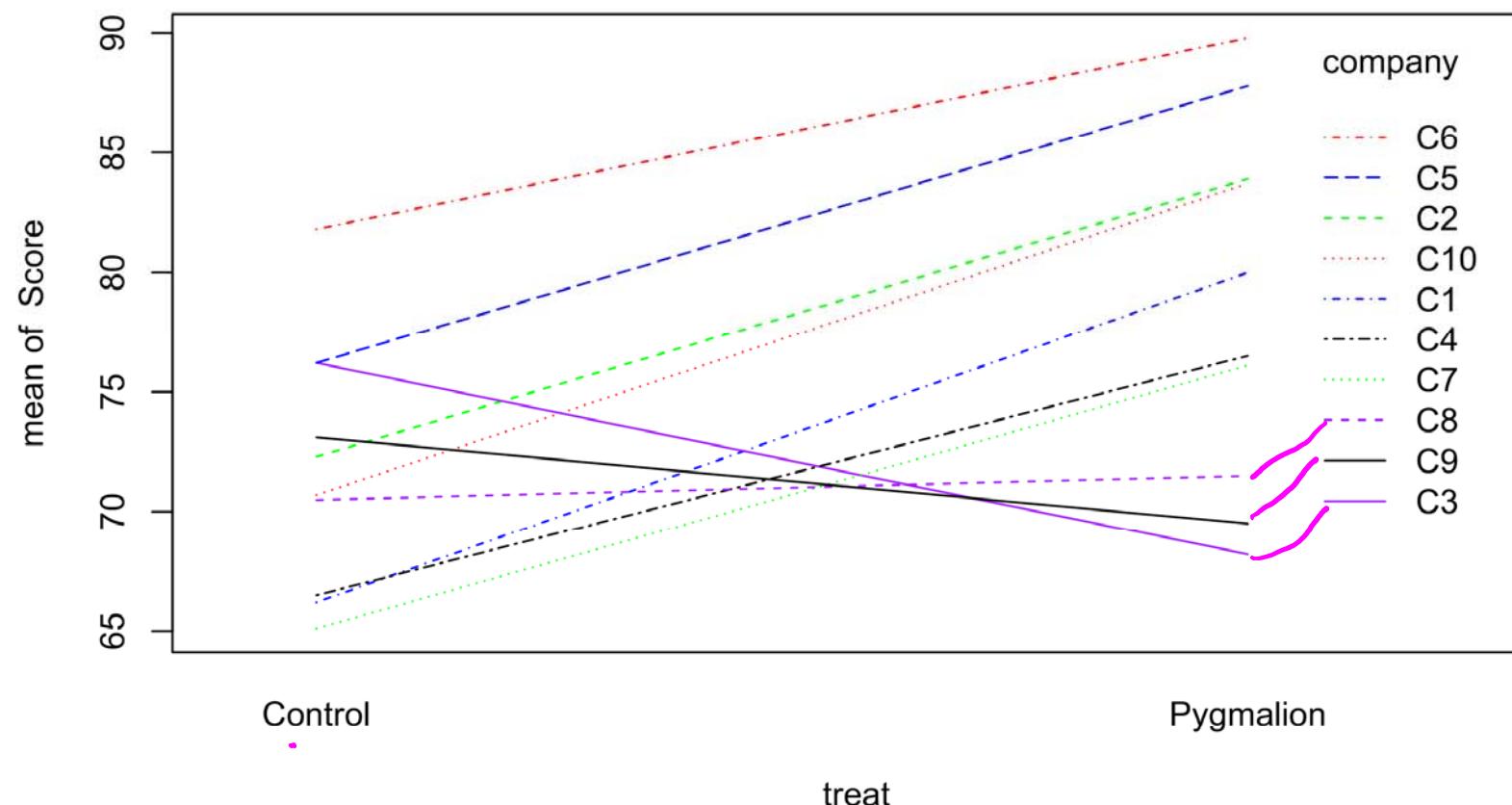
Case Study 2: Interaction plots

```
interaction.plot(company,treat,Score,col=c("blue","red"))
```



Case Study 2: Interaction plots

```
interaction.plot(treat,company,Score,col=c("blue", "red", "green","purple","bla
```



Case Study 2: Combination Means

```
cms=aggregate(Score~company+treat, data=pyg, FUN="mean")
cms[1:10,]
```

```
##      company   treat Score
## 1          C1 Control  66.2
## 2          C10 Control 70.7
## 3          C2 Control 72.3
## 4          C3 Control 76.2
## 5          C4 Control 66.5
## 6          C5 Control 76.2
## 7          C6 Control 81.8
## 8          C7 Control 65.1
## 9          C8 Control 70.5
## 10         C9 Control 73.1
```

Case Study 2: Combination Means

```
cms[11:20,]
```

```
##      company    treat Score
## 11       C1 Pygmalion  80.0
## 12      C10 Pygmalion  83.7
## 13       C2 Pygmalion  83.9
## 14       C3 Pygmalion  68.2
## 15       C4 Pygmalion  76.5
## 16       C5 Pygmalion  87.8
## 17       C6 Pygmalion  89.8
## 18       C7 Pygmalion  76.1
## 19       C8 Pygmalion  71.5
## 20       C9 Pygmalion  69.5
```

Case Study 2: Combination Means

```
tapply(Score, list(company,treat), mean)
```

```
##      Control Pygmalion
## C1      66.2      80.0
## C10     70.7      83.7
## C2      72.3      83.9
## C3      76.2      68.2
## C4      66.5      76.5
## C5      76.2      87.8
## C6      81.8      89.8
## C7      65.1      76.1
## C8      70.5      71.5
## C9      73.1      69.5
```

Case Study 2: Marginal Means

```
tapply(Score, company, mean)
```

```
##      C1      C10      C2      C3      C4      C5      C6      C7
## 70.80000 75.03333 76.16667 72.20000 69.83333 80.06667 84.46667 68.76667
##      C8      C9
## 70.83333 71.90000
```

```
tapply(Score, treat, mean)
```

```
##    Control Pygmalion
## 71.63158 78.70000
```

Case Study 2: Interaction model summary

```
##  
## Call:  
## lm(formula = Score ~ company * treat)  
##  
## Residuals:  
##    Min     1Q Median     3Q    Max  
##   -9.2   -2.3    0.0    2.3   9.2  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 66.200    5.094 12.996 3.89e-07 ***  
## companyC10                  4.500    7.204  0.625  0.5477  
## companyC2                  6.100    7.204  0.847  0.4191  
## companyC3                 10.000    8.823  1.133  0.2863  
## companyC4                  0.300    7.204  0.042  0.9677  
## companyC5                 10.000    7.204  1.388  0.1985  
## companyC6                 15.600    7.204  2.166  0.0585 .  
## companyC7                 -1.100    7.204 -0.153  0.8820  
## companyC8                  4.300    7.204  0.597  0.5653  
## companyC9                  6.900    7.204  0.958  0.3632  
## treatPygmalion              13.800    8.823  1.564  0.1522  
## companyC10:treatPygmalion -0.800   12.477 -0.064  0.9503  
## companyC2:treatPygmalion -2.200   12.477 -0.176  0.8639  
## companyC3:treatPygmalion -21.800  13.477 -1.618  0.1402  
## companyC4:treatPygmalion -3.800   12.477 -0.305  0.7676  
## companyC5:treatPygmalion -2.200   12.477 -0.176  0.8639
```

RSE = 7.204



Case Study 2: Interaction model

ANOVA Table

```
anova(lm(Score~company*treat))
```

FULL

```
## Analysis of Variance Table
##
## Response: Score
##              Df Sum Sq Mean Sq F value Pr(>F)
## company        9 670.98  74.55  1.4367 0.29902
## treat          2-1 = 1 338.88 338.88  6.5304 0.03092 *
## company:treat 9 311.46  34.61  0.6669 0.72212
## Residuals      9 467.04  51.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reg

Error

$SS_{Reg\ FULL}$

$10-1$

$$MSE = \hat{\sigma}^2 = (7.204)^2$$

RSS_{FULL}

$\frac{df}{G_1 - 1}$

$G_2 - 1$

$(G_1 - 1)(G_2 - 1)$

$N - G_1 G_2$

$N - 1$

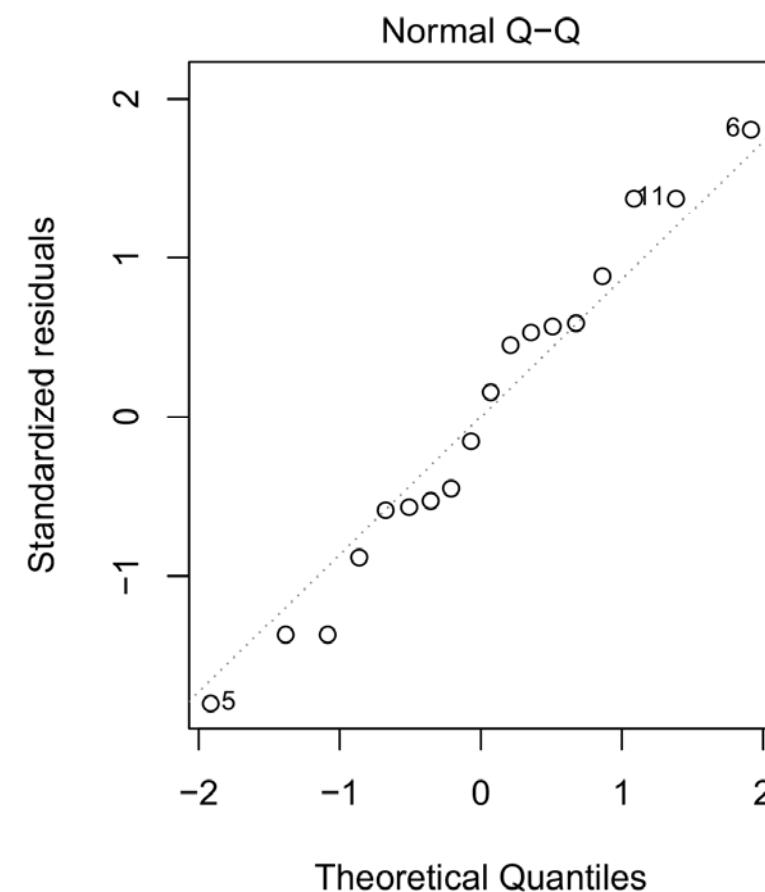
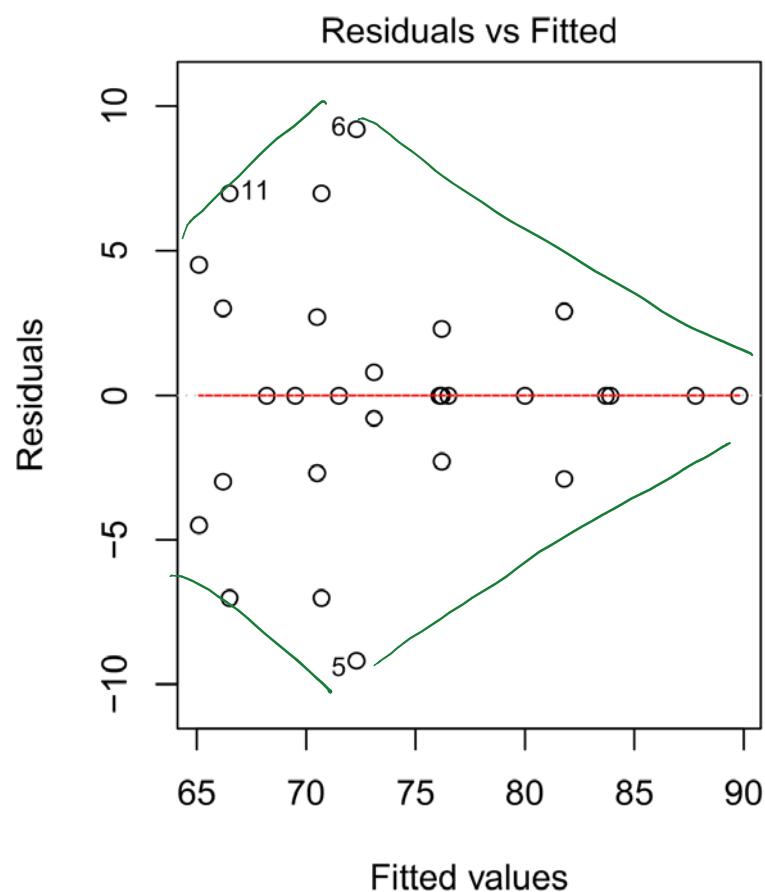
$$DF_{Total} = N - 1 = \# \text{ of observations} - 1 = 29 - 1 = 28$$

$$DF_{Error} = 28 - (19) = 9.$$

Case Study 2: Checking assumptions

```
fiti=lm(Score~company*treat, data=pyg)
par(mfrow=c(1,2))
plot(fiti, which=1:2)
```

```
## Warning: not plotting observations with leverage one:
##   1, 4, 7, 8, 9, 12, 15, 18, 21, 24, 27
```

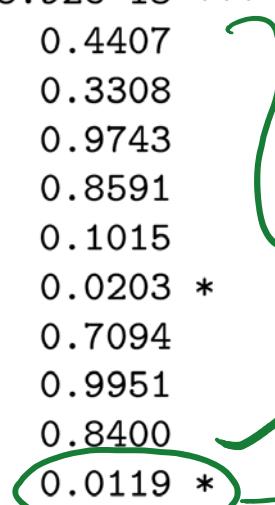


Case Study 2: Additive model summary

```
summary(lm(Score~company+treat))
```

```
##  
## Call:  
## lm(formula = Score ~ company + treat)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -10.660  -4.147   1.853   3.853   7.740  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 68.39316  3.89308 17.568 8.92e-13 ***  
## companyC10  4.23333  5.36968  0.788  0.4407  
## companyC2  5.36667  5.36968  0.999  0.3308  
## companyC3  0.19658  6.01886  0.033  0.9743  
## companyC4 -0.96667  5.36968 -0.180  0.8591  
## companyC5  9.26667  5.36968  1.726  0.1015  
## companyC6 13.66667  5.36968  2.545  0.0203 *  
## companyC7 -2.03333  5.36968 -0.379  0.7094  
## companyC8  0.03333  5.36968  0.006  0.9951  
## companyC9  1.10000  5.36968  0.205  0.8400  
## treatPygmalion 7.22051  2.57951  2.799  0.0119 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##
```

$$H_0: \beta_2 = 0$$



$$H_0: \beta_1 = 0$$

Case Study 2: Additive model

Reduced

```
anova(lm(Score~company+treat))  
## Analysis of Variance Table  
##  
## Response: Score  
##  
## Df Sum Sq Mean Sq F value Pr(>F)  
## company 9 670.98 74.55 1.7238 0.15556  
## treat 1 338.88 338.88 7.8354 0.01186 *  
## Residuals 18 778.50 43.25  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSReg reduced

RSS reduced

Red: $\text{Score} \sim \text{company}$
Full: $\text{Score} \sim \text{comp} + \text{treat}$

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$$
$$F = 7.8354$$
$$P\text{-value} = 0.012$$

Case Study 2: Additive model summary

```
summary(lm(Score~treat+company))
```

```
##  
## Call:  
## lm(formula = Score ~ treat + company)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -10.660  -4.147   1.853   3.853   7.740  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 68.39316  3.89308 17.568 8.92e-13 ***  
## treatPygmalion 7.22051  2.57951  2.799  0.0119 *  
## companyC10    4.23333  5.36968  0.788  0.4407  
## companyC2     5.36667  5.36968  0.999  0.3308  
## companyC3     0.19658  6.01886  0.033  0.9743  
## companyC4    -0.96667  5.36968 -0.180  0.8591  
## companyC5     9.26667  5.36968  1.726  0.1015  
## companyC6    13.66667  5.36968  2.545  0.0203 *  
## companyC7    -2.03333  5.36968 -0.379  0.7094  
## companyC8     0.03333  5.36968  0.006  0.9951  
## companyC9     1.10000  5.36968  0.205  0.8400  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##
```

Red: $\text{Score} \sim \text{treat}$
Full: $\text{Score} \sim \text{treat} + \underline{\text{company}}$

Case Study 2: Additive model

```
anova(lm(Score~treat+company))

## Analysis of Variance Table
##
## Response: Score
##             Df Sum Sq Mean Sq F value    Pr(>F)
## treat         1 327.34 327.34  7.5685 0.01314 *
## company       9 682.52   75.84  1.7534 0.14844
## Residuals  18 778.50    43.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H_0: \beta_2 = 0 = \beta_3 = \dots = \beta_{10}$$

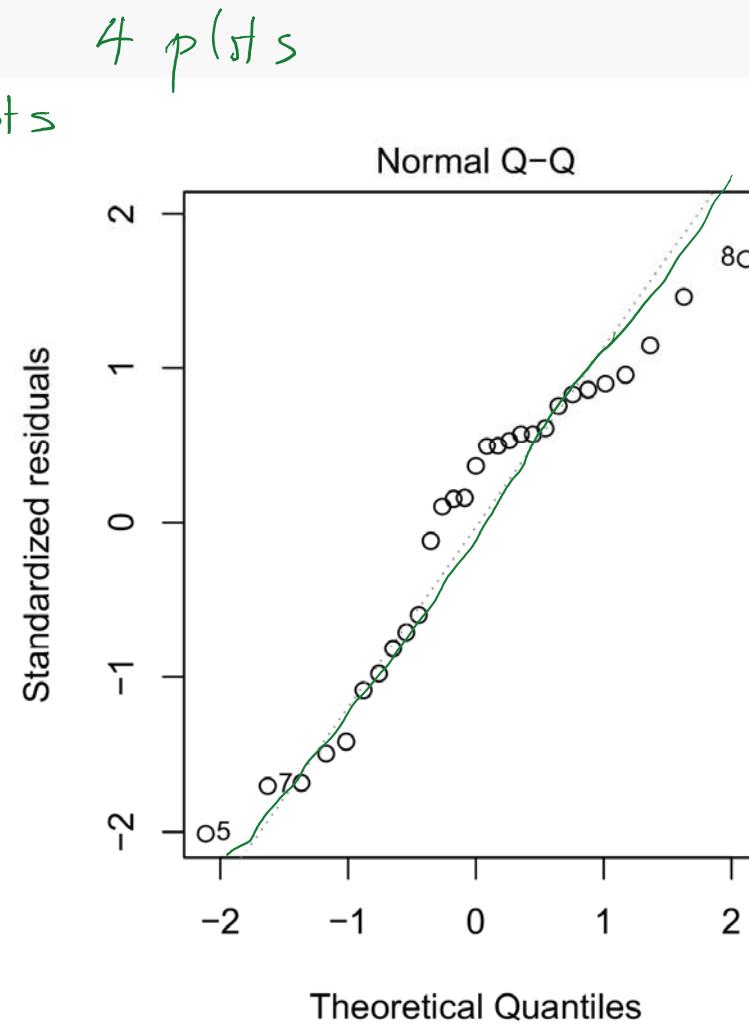
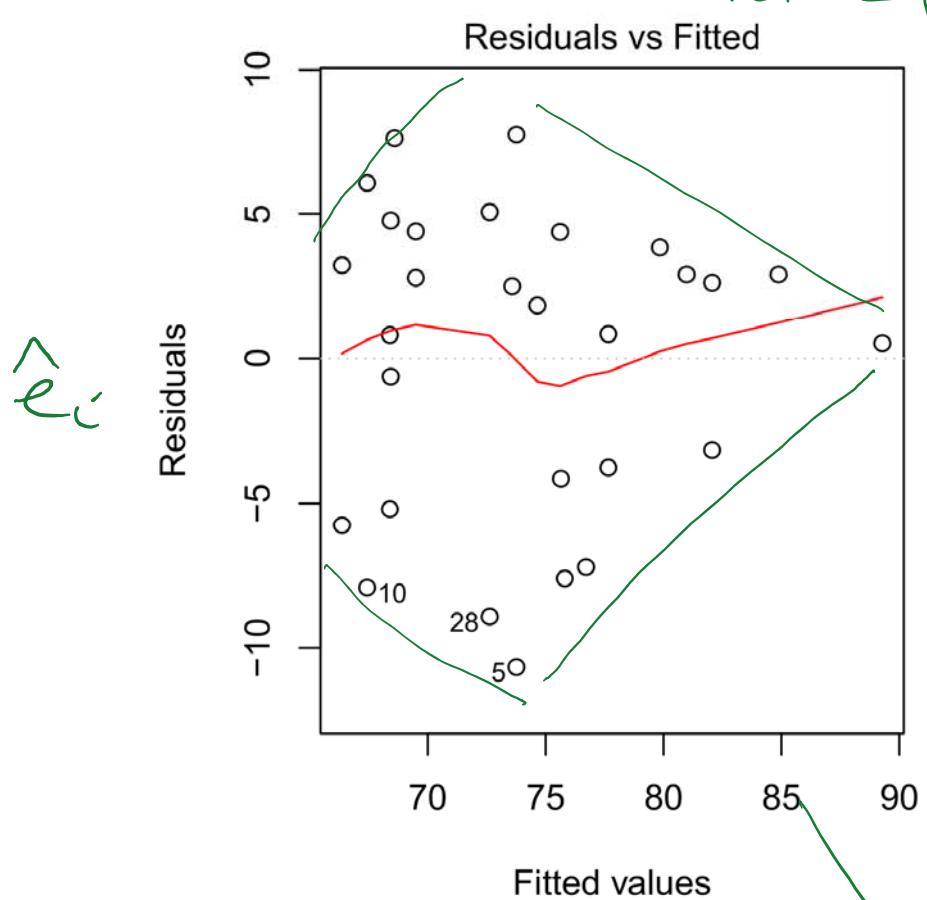
$$F = 1.75$$

$$P = 0.148 \text{ (large)}$$

Evidence of no diff
in companies.

Case Study 2: Checking assumptions

```
fita=lm(Score~company+treat, data=pyg)  
par(mfrow=c(1,2))  
plot(fita, which=1:2)
```



Possible decreasing variance \Rightarrow Weighted L-S.

4 plots

Normality satisfied

STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



January 23-25, 2018

Two-way ANOVA

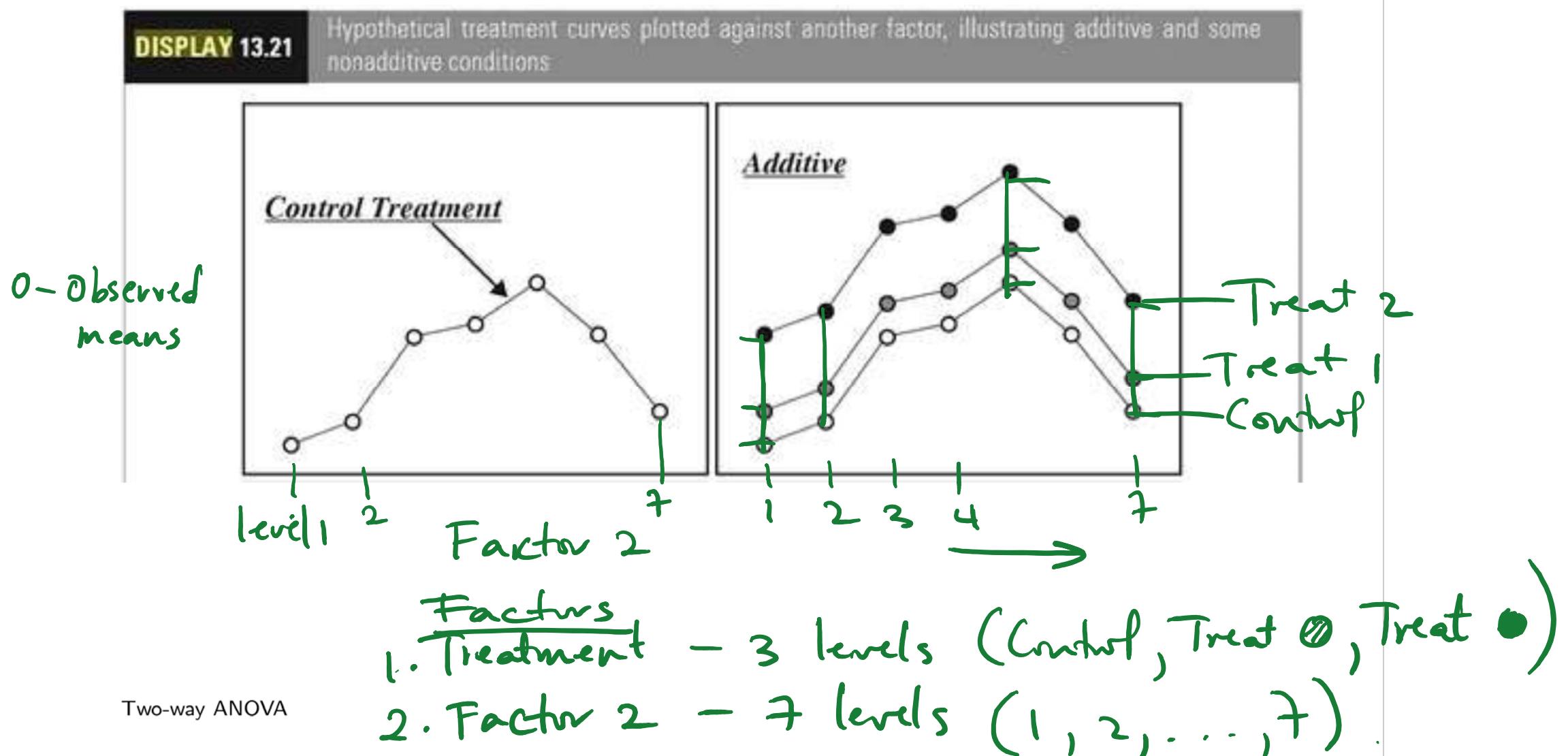
In the presence of “Interactions”

- ▶ Hard to answer questions about the main factor effects
- ▶ Communicate a table of estimated means , 
- ▶ Have separate models of Y against one factor for the different levels of the other factor

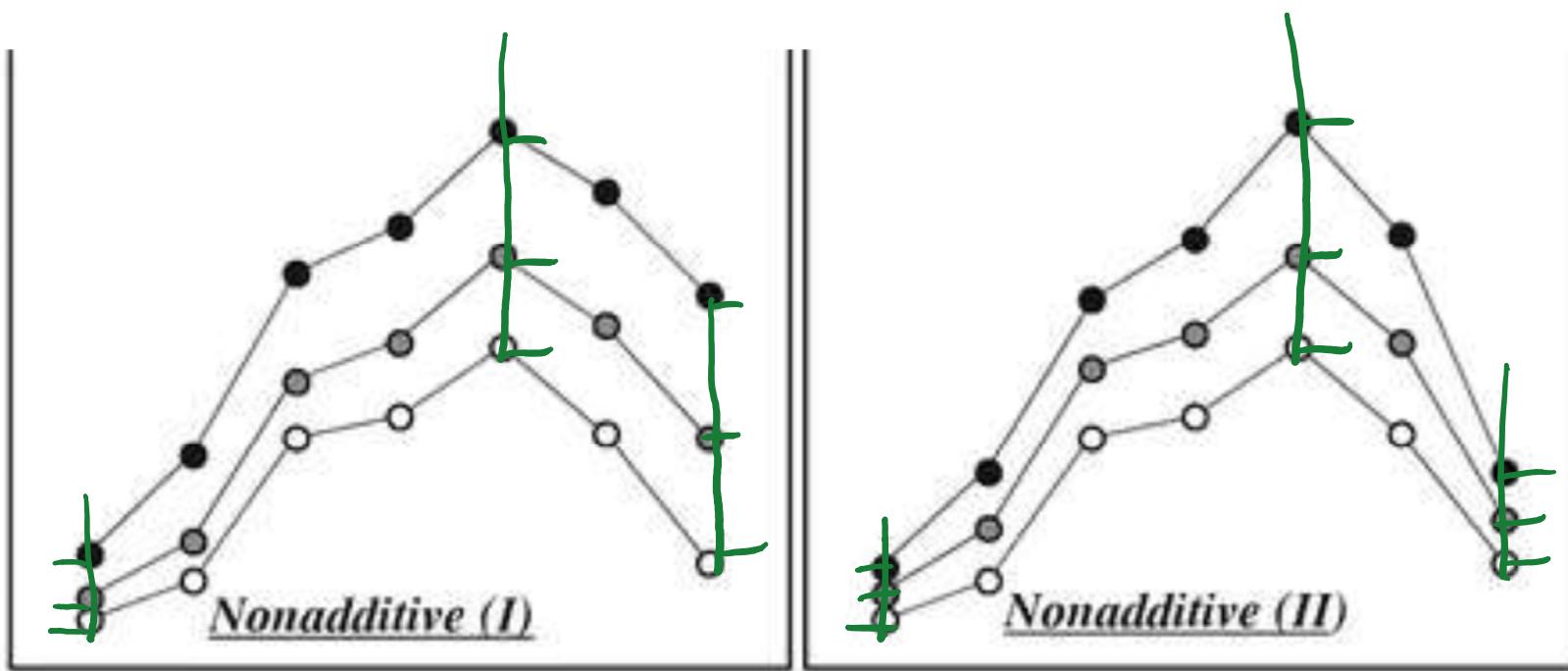
References:

- ▶ The Statistical Sleuth, 3rd edition by Ramsey and Schafer
- ▶ <https://cran.r-project.org/web/packages/Sleuth3/vignettes/chapter13-HortonMosaic.pdf>

In the presence of “Interactions”



In the presence of “Interactions”

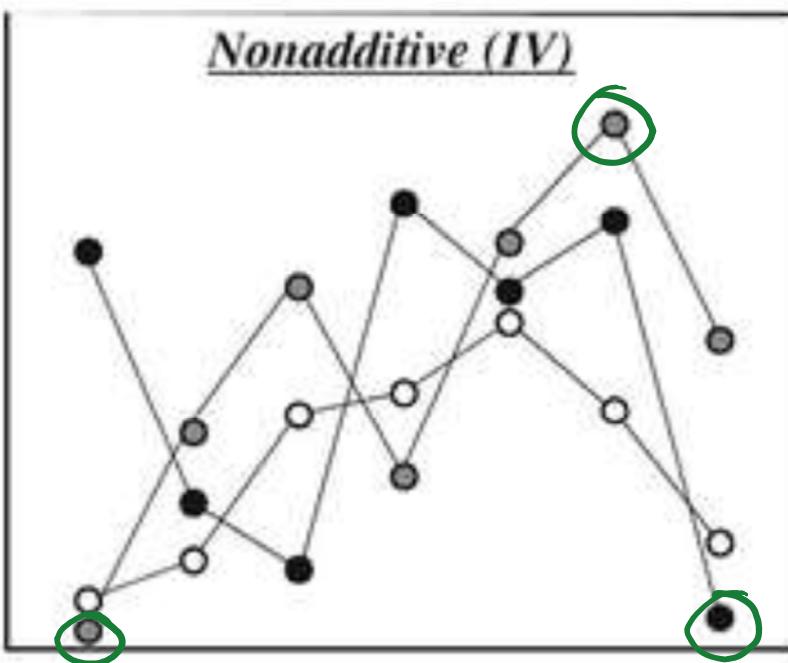
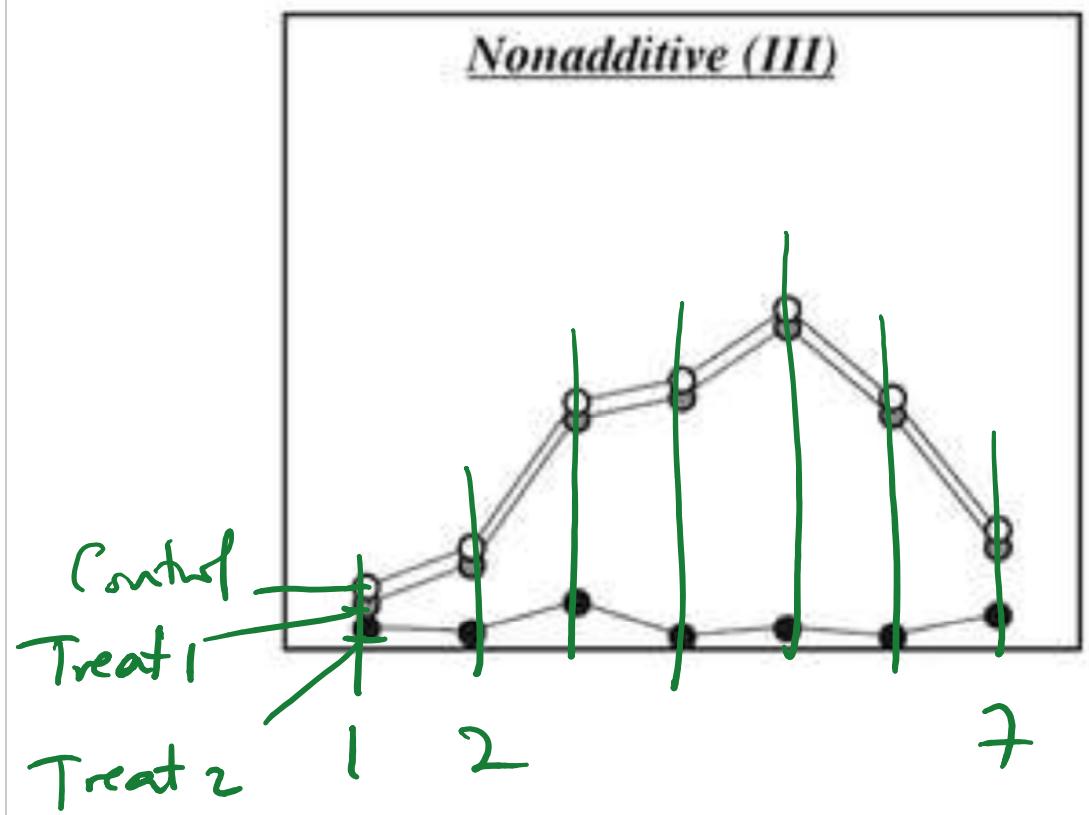


→
Systematic increase in
treatment effect
as we increase
level of Factor 2

Two-way ANOVA

For small observed
means, the treatment
effect is smaller
and vice versa

In the presence of “Interactions”



Two-way ANOVA
Treat 2's effect is different from that of Control or Treat 1 as we vary Factor 2's level.

Very complicated interaction
— Identify a few cases
— Add an additional factor

Should insignificant block effects be kept in the model?

- a / factn
- not treatment factn

- ▶ General advice is to drop insignificant terms
- ▶ For data from a randomized block experiment, block effects should be maintained
- ▶ Ensure that the control exercised by blocking is maintained in the analysis.

Case Study II-The Pygmalion Effect

- ▶ *Pygmalion effect*- high expectations of a supervisor or teacher translate to improved performance by subordinates or students
- ▶ Data:

Company	<u>Treatments</u>			
	Pygmalion	Control		
1	80.0	63.2	69.2	
2	83.9	63.1	81.5	
3	68.2	76.2		
4	76.5	59.5	73.5	
5	87.8	73.9	78.5	
6	89.8	78.9	84.7	
7	76.1	60.6	69.6	
8	71.5	67.8	73.2	
9	69.5	72.3	73.9	
10	83.7	63.7	77.7	

$$\frac{63.2 + 67.2}{2} = 65.2$$

→ 76.2

Case Study II: Additive model summary

Call:
lm(formula = Score ~ company + treat)

Coefficients: $\hat{\beta} \downarrow$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	68.39316	3.89308	17.568	8.92e-13 ***
companyC10	4.23333	5.36968	0.788	0.4407
companyC2	5.36667	5.36968	0.999	0.3308
companyC3	0.19658	6.01886	0.033	0.9743
companyC4	-0.96667	5.36968	-0.180	0.8591
companyC5	9.26667	5.36968	1.726	0.1015
companyC6	13.66667	5.36968	2.545	0.0203 *
companyC7	-2.03333	5.36968	-0.379	0.7094
companyC8	0.03333	5.36968	0.006	0.9951
companyC9	1.10000	5.36968	0.205	0.8400
treatPygmalion	7.22051	2.57951	2.799	0.0119 *

Residual standard error: 6.576 on 18 degrees of freedom
Multiple R-squared: 0.5647, Adjusted R-squared: 0.3228
F-statistic: 2.335 on 10 and 18 DF, p-value: 0.0564

Two-way ANOVA

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{10}$$

$$\begin{aligned}\hat{y} | (\text{Treat}, \text{Company}) \\ = \hat{\beta}_0 + \hat{\beta}_{\text{Treat}} + \hat{\beta}_{\text{Comp.}}\end{aligned}$$

Case Study II: Additive model summary

Call:
lm(formula = Score ~ treat + company)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	68.39316	3.89308	17.568	8.92e-13 ***
treatPygmalion	7.22051	2.57951	2.799	0.0119 *
companyC10	4.23333	5.36968	0.788	0.4407
companyC2	5.36667	5.36968	0.999	0.3308
companyC3	0.19658	6.01886	0.033	0.9743
companyC4	-0.96667	5.36968	-0.180	0.8591
companyC5	9.26667	5.36968	1.726	0.1015
companyC6	13.66667	5.36968	2.545	0.0203 *
companyC7	-2.03333	5.36968	-0.379	0.7094
companyC8	0.03333	5.36968	0.006	0.9951
companyC9	1.10000	5.36968	0.205	0.8400

Residual standard error: 6.576 on 18 degrees of freedom
Multiple R-squared: 0.5647, Adjusted R-squared: 0.3228
F-statistic: 2.335 on 10 and 18 DF, p-value: 0.0564

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_{10} = 0 \Rightarrow E(Y) = \beta_0$$

Two-way ANOVA

$$H_a: \text{at least } 1 \beta \neq 0 \quad (p=0.0564) < \alpha = 0.10$$

$$\hat{\beta}_i \pm t^* s_e(\hat{\beta}_i).$$

Estimated Mean Response from Additive Model

$$\hat{y}(\text{Treat}, \text{Comp}) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_{\text{Comp}}.$$

Company	Pygmalion ($\mathbb{1}_{PYG,i} = 1$)	Control ($\mathbb{1}_{PYG,i} = 0$)
1	$68.39 + 7.22 = \hat{\beta}_0 + \hat{\beta}_1$	$68.39 = \hat{\beta}_0$
2	$68.39 + 7.22 + 5.37 = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3$	$68.39 + 5.37$
3	.	.
4	.	.
5	.	.
6	.	.
7	.	.
8	.	.
9	.	.
10	$68.39 + 7.22 + 4.23$	$68.39 + 4.23 = \hat{\beta}_0 + \hat{\beta}_1$

$\downarrow \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$

Two-way ANOVA

Observed Group means vs Estimated means

Company	Observed Means			$n_{control}$	Estimated Means	
	Pyg	Control			Pyg	Control
1	80.0	66.2		2	75.61	68.39
2	83.9	72.3		2	80.98	73.76
3	68.2	76.2	1	2	75.81	68.59
4	76.5	66.5		2	74.65	67.43
5	87.8	76.2		2	84.88	77.66
6	89.8	81.8		2	89.28	82.06
7	76.1	65.1		2	73.58	66.36
8	71.5	70.5		2	75.65	68.43
9	69.5	73.1		2	76.71	69.49
10	83.7	70.7		2	79.85	72.63

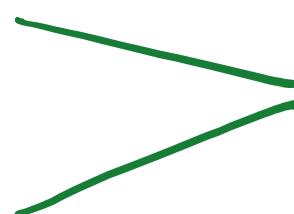
means 78.70 71.63 78.70 71.48

from 10 19 from 10 From 10 estimated means

Parameter estimation and Unbalanced design

- ▶ Estimated means for treatments are averages over 10 companies
- ▶ Observed Means vs Estimated means: Not the same because there are unequal number of control observations per company. Company 3 has 1 control platoon; other companies have 2.
- ▶ The design is nearly balanced.

- {
 - ▶ Affects constant variance assumption and variance estimate
 - ▶ Consider any evidence as exploratory
 - ▶ Consider *weighted* least squares regression



Measuring treatment effect

$$(\bar{x}_1 - \bar{x}_2) \pm t_{27, 0.025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

```
> qt(1-0.05/2, df=27)
> sqrt((9*var(Score[treat=="Pygmalion"])+18*var(Score[treat=="Control"]))/27)
> t.test(Score[treat=="Pygmalion"], Score[treat=="Control"], var.equal=T)
```

[1] 2.051831 ↑
[1] 7.356078 ↑

Two Sample t-test

$$n_1 + n_2 - 2 = 10 + 19 - 2 = 27.$$

```
data: Score[treat == "Pygmalion"] and Score[treat == "Control"]
t = 2.4595, df = 27, p-value = 0.0206
alternative hypothesis: true difference in means is not equal to 0
```

95 percent confidence interval
1.171707 12.965135

sample estimates:
mean of x mean of y
78.70000 71.63158

← does not include 0 .

Two-way ANOVA

Conclusions

- ▶ There is evidence of a difference in mean score between pygmalion and control platoons ($p=0.0119$). (Consider this as weak evidence since we have some concerns about variance estimates.)
- ▶ Confidence Intervals for the difference in mean score between pygmalion and control platoons:

- ▶ Pooled 2-sample t:

$$(78.7 - 71.6) \pm 2.05(7.36)\sqrt{(1/10 + 1/19)} = (1.17, 12.96)$$

- ▶ Least-squares approach (Additive model):

$$18 = 10 + 10 - 2 = df$$

$\hat{\beta}_1$ → $7.22 \pm 2.101(2.5795) = (1.8, 12.6)$ $se(\hat{\beta}_1)$

Similar
since
design
is nearly
balanced.

$$\begin{aligned} 78.7 &- 71.6 \\ = 7.22 \end{aligned}$$

- ▶ On average, pygmalion platoons (mean=78.7) scored higher than control platoons (mean=71.6).

STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



January 30- February 1, 2018

1/34

STA 303/1002: Class 8-Generalized Linear Models

- ▶ Case Study III: The Donner Party Example
- ▶ Generalized Linear Models
 - ▶ What is a Generalized Linear Model?
 - ▶ Common link functions
 - ▶ What is a Binary Logistic Regression Model?
 - ▶ Maximum likelihood estimation of β 's
- ▶ Case Study III Example
 - ▶ Data and Questions
 - ▶ Estimated Model
 - ▶ Interpretations

Case Study III: The Donner Party Example

- ▶ Background: (D.K. Grayson, Journal of Anthropological Research, 1990: 223-42)
 - ▶ In mid 19th century, a group of 86 American pioneers headed out from Missouri toward California in a wagon train.
 - ▶ Due to a combination of harsh weather, unsuitable travel equipment and divisions with the party, the group got stuck in the Sierra Nevada mountain range.
 - ▶ They had planned to arrive safe and sound in September but those who survived did not make it there until the following March.
- ▶ Question: Who survived?- Men? -Older pioneers?
- ▶ Data:
 - ▶ age
 - ▶ sex
 - ▶ outcome: survived or not
- ▶ AIM: Study the odds of survival

Ramsey &
Schafer, 3rd ed.

Case Study III: Model

"success", $y=1$
"failure", $y=0$

- ▶ Response: Y_i - a binary variable (eg., survived or died)
- ▶ Predictor: X_i - eg., age, sex of i th pioneer
- ▶ Model: BINARY LOGISTIC REGRESSION

$$Y_i|X_i = \begin{cases} 1 & \text{if response is in category of interest} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_i|X_i \sim \text{Bernoulli}(\pi_i)$$

Then:

- ▶ $E[Y_i|X_i] = \pi_i$ and $\text{Var}(Y_i|X_i) = \pi_i(1 - \pi_i)$
- ▶ A logistic regression model is an example of a **Generalized Linear Model**.

$$\epsilon_i = y_i - \hat{y}_i$$

Obs / *Exp.*

Generalized Linear Models

- ▶ Have:
 - response, Y and
 - a set of explanatory variables X_1, \dots, X_p
- ▶ Want: Model $E(Y)$ as a **linear** function in the parameters, ie.,

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = \mathbf{X}\boldsymbol{\beta}$$

- ▶ Key idea: Choice of the **link function**, g such that

$$g(E(Y)) = \mathbf{X}\boldsymbol{\beta}$$

$E(g(Y)) \leftarrow$ ^{5/34}transform Y

Some Link Functions

Let $E(Y) = \mu$.

General →

Link	Function	Usual distribution of $Y X$	$\epsilon_i \sim N(0, \sigma^2 I)$
Identity	$g(\mu) = \mu$	Normal	
Log	$g(\mu) = \log \mu, \mu > 0$	Poisson (count data)	
Logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right), 0 < \mu < 1$	Bernoulli (binary), Binomial	

Note: Link function, $g(\cdot)$ is a function of $\mu = E(Y)$, the mean of Y , and not a transformation of the data.

$$\begin{aligned} &\log(E(y)) \\ &g(E(y)) \end{aligned}$$

$$\begin{aligned} &E \log(y) \\ &E[g(y)] \end{aligned}$$

Binary Logistic Regression

GLMs vs Transforming the data

- ▶ Transform Y so it has an approximate normal distribution with constant variance. Common variance stabilizing transformations (Weisberg, 3rd ed, p. 179):
 - ▶ \sqrt{Y} : mild transformation; used when $Var(Y|X) \propto E(Y|X)$ as for Poisson data
 - ▶  $\log(Y)$: most common; if $Var(Y|X) \propto [E(Y|X)]^2$ or errors behave like percentage of Y .
 - ▶ $1/Y$: used when responses are mostly close to 0, but some large values occur.
- ▶ As GLM (Agresti, p. 117):
 - ▶  distribution of Y not restricted to Normal
 - ▶ model parameters describe $g[E(Y)]$ rather than $E(g(Y))$ as in transformed data approach
 - ▶ GLMs provide a unified theory of modelling that encompasses the most important models for continuous and discrete variables.

7/34

LOG ODDS, ODDS, ODDS RATIO

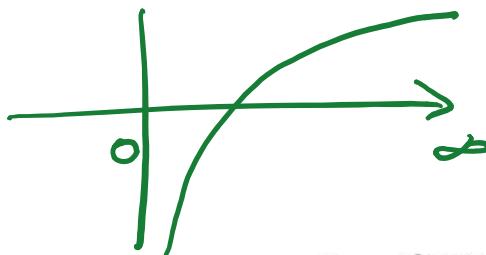
- ▶ Let $\pi = P(\text{"success"})$, $0 < \pi < 1$.
- ▶ The ODDS in favour of "success" is:

$$\frac{\pi}{1 - \pi}$$

- ▶ Then the LOG ODDS is: $(-\infty, \infty)$

$$\log\left(\frac{\pi}{1 - \pi}\right)$$

- ▶ An ODDS RATIO is a ratio of ODDS.



$$\begin{aligned} l \\ (0, \infty) \\ \frac{P(\text{"success"})}{P(\text{"failure"})} \\ \text{As } \pi \rightarrow 0, \text{ ODDS} \rightarrow 0 \\ \text{As } \pi \rightarrow 1, \text{ ODDS} \rightarrow \infty \\ (0, \infty) \end{aligned}$$

Binary Logistic Regression

- ▶ $E(Y|X) = \pi$
- ▶ $\text{Var}(Y|X) = \pi(1 - \pi)$. Notice that variance is not constant!
- ▶ Logistic regression model:

$$\log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (1)$$

- ▶ Linear predictor:

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- ▶ LOGISTIC FUNCTION: Find by inverting equation (1)

$$\underline{\pi(\eta)} = \frac{e^M}{1 + e^M}$$

$$\log \frac{\pi}{1 - \pi} = M$$

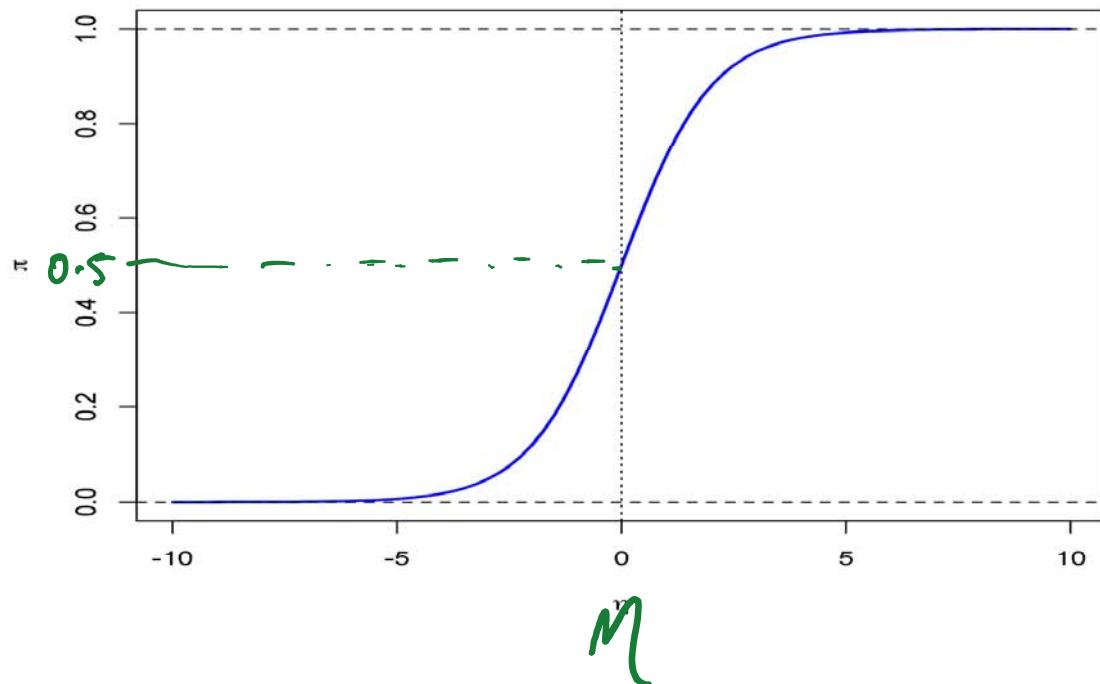
9/34

$$\frac{\pi}{1 - \pi} = e^M$$

$$\pi = \frac{e^M}{1 + e^M}$$

What does the logistic function look like?

- ▶ LOGISTIC FUNCTION: $\pi = \frac{e^\eta}{1+e^\eta}$



- ▶ S-shaped; sigmoid function
- ▶ Horizontal asymptotes at 0 and 1; the logistic function, $\pi(\eta)$ varies between 0 and 1

10/34

Binary Logistic Regression Model

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \quad i = 1, \dots, n$$

- ▶ Log-odds, $\log(\pi/(1 - \pi))$ are between $-\infty$ and ∞ (good characteristic of a link function)
- ▶ As π_i (the probability of “success”) increases, odds of success and log-odds increase
- ▶ Predicts the natural log of the odds for a subject being in one category or another
- ▶ Regression coefficients can be used to estimate odds ratio for each of the independent variables
- ▶ Tells which predictors can be used to determine if a subject was in a category of interest

How to estimate the parameter coefficients?

Maximum Likelihood Estimation

- ▶ Data: $Y_i = \begin{cases} 1 & \text{if response is in category of interest} \\ 0 & \text{otherwise} \end{cases}$
- ▶ Model: $P(Y_i = y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$
- ▶ Assume: The n observations are independent
- ▶ Joint density:

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \pi_1^{y_1} (1-\pi_1)^{1-y_1} \times \dots \times \pi_n^{y_n} (1-\pi_n)^{1-y_n}$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})} = \frac{e^{M_i}}{1 + e^{M_i}}$$

$$\text{and } 1 - \pi_i = 1 - \frac{e^{M_i}}{1 + e^{M_i}} = \frac{1 + e^{M_i} - e^{M_i}}{1 + e^{M_i}} = \frac{1}{1 + e^{M_i}}$$

12/34

Maximum Likelihood Estimation

- ▶ **Likelihood function:** Plug in observed data and think of the joint density as a function of β 's-

$$\mathcal{L}(\beta_0, \dots, \beta_p) = \prod_{i=1}^n \pi_i(\beta)^{y_i} (1 - \pi_i(\beta))^{1-y_i}$$
$$= \prod_{i=1}^n \left(\frac{e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}} \right)^{1-y_i}$$

$$\log \mathcal{L}(\beta_0, \dots, \beta_p) = \sum_{i=1}^n [y_i (\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}) - y_i \log(1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})) - (1 - y_i) \log(1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}))]$$

- ▶ Maximize the log-likelihood:

$$(\hat{\beta}_0, \dots, \hat{\beta}_p) = \arg \max \{\log \mathcal{L}(\beta_0, \dots, \beta_p)\}$$

13/34

$$\log AB = \log A + \log B.$$

$$\log C^D = D \log C$$

$$\log \frac{1}{a} = -\log a$$

$$\log a^{-1} = -\log a$$

MLE solution methods

- ▶ No explicit expression exists for the maximum likelihood estimators $-(\hat{\beta}_0, \dots, \hat{\beta}_p)$.
- ▶ Two iterative numerical solution methods are:
 - (1) Newton-Raphson algorithm
 - (2) Fisher scoring or Iteratively Re-weighted Least Squares (IWLS). This is done in glm().

Large-sample properties of MLEs

If model is correct, and sample size is large enough, as $n \rightarrow \infty$

1. MLEs are unbiased
2. MLEs have minimum variance
3. MLEs are Normally distributed
4. Formulas for standard errors of MLEs are well-known.

Estimates of standard errors are available as by-product of numerical optimization (maximization) procedures.

$$\hat{\beta}_{MLE} \pm z_{\alpha/2} se(\hat{\beta}_{MLE})$$

Case Study III: The Donner Party Example

16/34

Case Study III: The Data

- ▶ Data: $n=45$ pioneers

AGE	SEX	STATUS
23	MALE	DIED
40	FEMALE	SURVIVED
40	MALE	SURVIVED
30	MALE	DIED
28	MALE	DIED
40	MALE	DIED
...		

- ▶ AGE: Adults, 15-65 yrs old
- ▶ SEX: 15 Females, 30 Males
- ▶ BINARY OUTCOME: 25 Died, 20 Survived
- ▶ Questions: What are the odds of survival for a 20-yr old female? Compare the odds of survival to that of a male of the same age.

17/34

Case Study III: Binary Logistic Regression Additive Model

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 \text{Age}_{i1} + \beta_2 \text{Sex}_{i2}, \quad i = 1, \dots, 45$$

- ▶ Cannot predict survival ($\pi = 1$) or death ($\pi = 0$) of a pioneer
- ▶ Can estimate:
 - ▶ π_i (the probability of survival)
 - ▶ odds of survival and
 - ▶ log-odds of survival based on *Age* and *Sex* of a pioneer
- ▶ Can be used to get point and interval estimates of odds ratios
- ▶ Can test which predictors are relevant to determine odds of survival

Interpreting coefficients of a Binary Logistic model

For $\pi = P(Y = 1)$, we model

$$\log \left(\frac{\pi}{1 - \pi} \right) = \text{log odds}_{\{Y=1\}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Let ω be the odds that $Y=1$ based on X_1, \dots, X_p , then

$$\omega = \exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}.$$

Interpretation of β_1 : Holding $\underline{X_2}, \dots, \underline{X_p}$ fixed, the ratio of the odds ('ODDS RATIO') that $Y=1$ at $\underline{X_1=a}$ to $\underline{X_1=b}$ is

$$\frac{\omega_a}{\omega_b} = \exp\{\beta_1(a - b)\}.$$

If X_1 increases by 1 unit, holding all other X 's constant, the odds that $Y=1$ change by a multiplicative factor of e^{β_1} . 19/34

$$\frac{\omega_a}{\omega_b} = \frac{e^{\beta_0 + \beta_1 a + \beta_2 x_2 + \cdots + \beta_p x_p}}{e^{\beta_0 + \beta_1 b + \beta_2 x_2 + \cdots + \beta_p x_p}}$$

Using R for fitting GLMs

$y \sim X$

- ▶ fitting function:

`glm(formula, family, data)`

- ▶ family: link function, distribution of Y.

Examples include binomial, gaussian, poisson, Gamma

- ▶ complementary functions:

▶ `coefficients()`: coefficient estimates

▶ `summary()`: prints a summary of results

▶ `anova()`: produces an analysis of variance table

{
▶ `residuals`
▶ `deviance`

- ▶ Optimization technique: Fisher Scoring / IWLS

Case Study 3: The Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case2001
library(Sleuth3)
#Donner party survival data
donner = case2001
str(donner)

## 'data.frame': 45 obs. of 3 variables:
## $ Age : int 23 40 40 30 28 40 45 62 65 45 ...
## $ Sex : Factor w/ 2 levels "Female","Male": 2 1 2 2 2 2 1 2 2 1 ...
## $ Status: Factor w/ 2 levels "Died","Survived": 1 2 2 1 1 1 1 1 1 1 ...

attach(donner)
head(donner)

##   Age   Sex   Status
## 1 23 Male Died
## 2 40 Female Survived
## 3 40 Male Survived
## 4 30 Male Died
## 5 28 Male Died
## 6 40 Male Died
```

ref. grp

Case Study 3: Summarizing the data

```
#two-way contingency table for status by sex  
#check that cell counts>0  
xtabs(~Status+Sex, data=donner)
```

```
##           Sex  
## Status     Female Male  
##   Died        5   20  
##   Survived    10  10
```

10 Males survived

```
summary(Age)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.  
##      15.0    24.0    28.0    31.8    40.0    65.0
```

Case Study 3: Marginal Mean Ages

```
tapply(Age, Status, mean)
```

```
##      Died Survived  
##    35.48   27.20
```

avr. age by status

```
tapply(Age, Sex, mean)
```

Av. age by sex

```
## Female     Male  
## 31.06667 32.16667
```

```
fita<-glm(Status~Age+Sex, family=binomial, data=donner)
```

y | x_1 | x_2 | logit link

Case Study 2: Additive model summary

```
##          Summary(fit)
## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
## 
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.7445 -1.0441 -0.3029  0.8877  2.0472
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.23041   1.38686  2.329   0.0198 *
## Age        -0.07820   0.03728 -2.097   0.0359 *
## SexMale    -1.59729   0.75547 -2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
## 
## Number of Fisher Scoring iterations: 4
```

$\hat{\pi} = P(\text{"survival"})$

$X_2 = 1 \text{ for Male}$

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

$\hat{\beta}$

Case Study 3: ANOVA table

```
anova(fita)

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev
## NULL              44     61.827
## Age    1    5.5358      43     56.291
## Sex    1    5.0344      42     51.256
```

Case Study 3: Modelling "Died"

```
status=relevel(Status, ref="Survived")
fitad<-glm(status~Age+Sex, family=binomial, data=donner)
summary(fitad)
```

```
##
## Call:
## glm(formula = status ~ Age + Sex, family = binomial, data = donner)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.0472 -0.8877  0.3029  1.0441  1.7445
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.23041   1.38686 -2.329   0.0198 *
## Age          0.07820   0.03728  2.097   0.0359 *
## SexMale      1.59729   0.75547  2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Dispersion parameter for binomial family taken to be 1
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
```

$$\pi = P(\text{"died"})$$

$$\log \left(\frac{P(\text{"Survived"})}{P(\text{"died"})} \right)$$

$$\log \frac{P(\text{"died"})}{P(\text{"Survived"})}$$

$$\log \frac{a}{b} = \log a - \log b$$

$$\log \frac{b}{a} = \log b - \log a$$

Case Study 3: Sex Reference group as “Male”

```
sex=relevel(Sex, ref="Male")
fitadf<-glm(status~Age+sex, family=binomial, data=donner)
summary(fitadf)

##
## Call:
## glm(formula = status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.0472  -0.8877   0.3029   1.0441   1.7445
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.63312  1.11018 -1.471   0.1413
## Age          0.07820  0.03728  2.097   0.0359 *
## sexFemale   -1.59729  0.75547 -2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
```

$$\pi = P(\text{died})$$

$$x_2 = 1 \text{ if Female}$$

Case Study 3: Sex Reference group as “Male”

```
fitasf<-glm(Status~Age+sex, family=binomial, data=donner)
summary(fitasf)
```

```
##
## Call:
## glm(formula = Status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.7445   -1.0441   -0.3029    0.8877    2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471   0.1413
## Age         -0.07820   0.03728  -2.097   0.0359 *
## sexFemale   1.59729   0.75547   2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

$$\pi = P(\text{Survived})$$

$$\chi_2 = 1 \text{ if Female}$$

Case Study III: Fitted equations

Using *defaults*, $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = 3.23 - 0.078Age_i - 1.60\mathbb{1}_{Male,i}$$

Using other reference status, $\pi = P(DIED)$:

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = -3.23 + 0.078Age_i + 1.60\mathbb{1}_{Male,i}$$

Using sex reference group as Males, $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = 1.63 - 0.078Age_i + 1.60\mathbb{1}_{Female,i}$$

Using sex reference group as Males, $\pi = P(DIED)$:

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = -1.63 + 0.078Age_i - 1.60\mathbb{1}_{Female,29/34,i}$$

Case Study III: Using Fitted equation

Using the fitted equation for $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = 1.63 - 0.078Age_i + 1.60\mathbb{1}_{Female,i},$$

Q: Estimate the log odds, odds and probability of survival for a:

	Log odds, $\log(\frac{\hat{\pi}}{1-\hat{\pi}})$	Odds, $\frac{\hat{\pi}}{1-\hat{\pi}}$	$\hat{\pi} = \frac{\text{Odds}}{1+\text{Odds}}$
(i) 20-yr old Female	$1.63 - 0.078(20) + 1.6$		
(ii) 40-yr old Female	$1.63 - 0.078(40) + 1.6$		
(iii) 20-yr old Male	$1.63 - 0.078(20)$		
(iv) 40-yr old Male	$1.63 - 0.078(40)$	$e^{(1.63 - 0.078(40))}$	

Case Study III: Using Fitted equation

Using the fitted equation for $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = 1.63 - 0.078Age_i + 1.60\mathbb{1}_{Female,i},$$

Q: Estimate the log odds, odds and probability of survival for a:

	Log odds, $\log(\frac{\hat{\pi}}{1-\hat{\pi}})$	Odds, $\frac{\hat{\pi}}{1-\hat{\pi}}$	$\hat{\pi}$	
(i) 20-yr old Female	1.67	5.31	0.84	$\hat{\pi} = \frac{5.31}{1+5.31}$
(ii) 40-yr old Female	0.11	1.12	0.53	
(iii) 20-yr old Male	0.07	1.07	0.52	
(iv) 40-yr old Male	-1.49	0.225	0.18	

Qs: Compare the odds of survival for a 40-yr old Female to that of a 20-yr old Female. Compare the odds of survival for a 20-yr old Female to that of a Male of the same age.

31/34

Case Study III: Using coefficients to find Odds Ratios

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 = 1.63 - 0.078 \text{Age} + 1.60 \mathbb{1}_{Female}$$

1. (Fixed Sex) Compare the odds of survival for a 40-yr old (Female/Male) to that of a 20-yr old (Female/Male).

$$\exp\{-0.78(40 - 20)\} = 0.21 \approx \frac{1}{5}$$

Hence, the odds of survival for a 20-yr old are about 5 times the odds for a 40-yr old of the same sex.

2. (Fixed Age) Compare the odds of survival for a 20-yr old Female to that of a Male of the same age.

$$\exp\{1.60(1 - 0)\} = 4.95 \approx 5$$

Hence, the odds of survival for a Female are about 5 times the odds for a Male of the same age.

32/34

- nutrition
- social norm

Case Study III: Odds Ratios

1. Compare the odds of survival for a 40-yr old Female to that of a 20-yr old Female.

$$\frac{1.12}{5.31} = 0.21 \approx \frac{1}{5}$$

See page 31.

Hence, the odds of survival for a 20-yr old Female are about 5 times the odds for a 40-yr old Female.

2. Compare the odds of survival for a 20-yr old Female to that of a Male of the same age.

$$\frac{5.31}{1.07} = 4.96 \approx 5$$

Hence, the odds of survival for a 20-yr old Female are about 5 times the odds for a Male of the same age.

Next Class

- ▶ Confidence interval for Odds Ratio
- ▶ Testing β 's → Higher-order Models

STA 303/1002: Class 9- Case Study III Inference

Binary Logistic Regression Example

- ▶ Case Study III: The Donner Party Example
 - ▶ Confidence interval for Odds Ratio
 - ▶ Testing β 's → Higher-order Models
- ▶ Joke: "*I asked a statistician for her phone number... and she gave me an estimate.*"(www.workjoke.com)

Logistic Regression: Testing whether single β 's are zero

WALD CHI-SQUARE PROCEDURES

- ▶ **Hypotheses:** $H_0 : \beta_j = 0$ (X_j has no effect on log-odds)
 $H_a : \beta_j \neq 0$

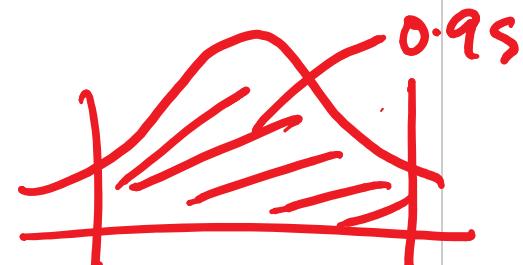
- ▶ **Test Statistic:**
$$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

where

- ▶ $\hat{\beta}_j$ - maximum likelihood (ML) estimate and
- ▶ $SE(\hat{\beta}_j)$ - estimated standard error from the numerical procedure that generated the MLE.
- ▶ By standard large-sample results, MLE's are normally distributed. Thus, for large n , under H_0 , z is an observation from an approx. $\mathcal{N}(0, 1)$ distribution.
- ▶ **95% Confidence interval:**
$$\hat{\beta}_j \pm 1.96SE(\hat{\beta}_j)$$

$$\begin{aligned} \text{logit}(x) &= \mathbf{x} \boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_1 + \dots \end{aligned}$$

$$\text{Est} \pm z_{\alpha/2} SE(\text{Est}) .$$



Examples: Testing whether single β 's are zero

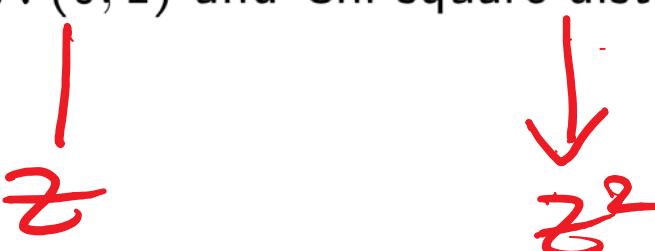
Using R output ('Coefficients'):

	Age	Sex
Test statistic	$(-0.078/0.0373)^2$	
P-value	0.036	
95% CI for β	$-0.078 \pm 1.96(0.0373)$ $=(-0.15, -0.0055)$	
CI for Odds ratio	$(e^{-0.15}, e^{-0.0055}) = (0.86, 0.995)$	
Conclusion	For the same sex, the odds ratio for a 1-year increase in age is between .86 and 0.995.	

In R Output

$$\hat{\beta}_j \pm z_{\alpha/2} \text{SE}(\hat{\beta}_j)$$

Recall relationship between $\mathcal{N}(0, 1)$ and Chi-square distribution:

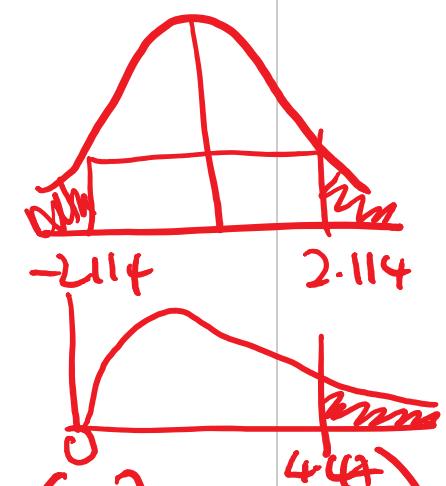


Examples: Testing whether β 's are zero

Using R output:

	Age	Sex
Test statistic	$(-0.078/0.0373)^2$	4.47 = 2.114 (Refer to Additive model)
P-value	0.036	0.0345
95% CI for β	$-0.078 \pm 1.96(0.0373)$ =(-0.15, -0.0055)	(0.117, 3.078)
CI for Odds ratio	$(e^{-0.15}, e^{-0.0055}) = (0.86, 0.995)$	(1.124, 21.72)
Conclusion	For the same sex, the odds ratio for a 1-year increase in age is between .86 and 0.995.	

- ▶ Note: Both marginal p-values are less than 0.05 and the confidence intervals for the odds ratios do not include 1.
- ▶ Hence, we have moderate evidence that both Age and Sex have an effect on survival over and above each other.
- ▶ Recall: If $Z \sim \mathcal{N}(0, 1)$, then $Z^2 \sim \chi_1$.



Binary Logistic Regression

$$\text{P-value} = 2 \cdot (1 - \text{pnorm}(2.114)) = P(\chi_1^2 > 4.47).$$

In R: $2(1 - \text{pnorm}(2.114))$, $1 - \text{pchisq}(4.47, df=1)$

Additional CI Examples

Using R output:

- ▶ Q: Find a 95% CI for the change in odds of survival for a 40-yr old to 20-yr old of the same sex.
- ▶ A:
 - ▶ The log odds change by $-0.078*(40-20) = -1.56$.
 - ▶ 95% CI for the change in log odds is $20 * (-0.15, -0.0055)$
 $= (-3.0, -0.11)$.
An arrow points from the word "log" in the original sentence to the "log" in this equation.
 - ▶ 95% CI for the odds ratio is $(0.05, 0.896)$. $= (e^{-3.0}, e^{-0.11})$
 - ▶ The odds of survival of a 40-yr old woman were $e^{-1.56} = 0.21$ times the odds of survival for a 20-yr old.
- ▶ Note that it is not appropriate to compute CI for π since $0 \leq \pi \leq 1$ and it is not normally distributed.

Model Assumptions for Binary Logistic Regression

1. Underlying probability model for response is Bernoulli.
2. Observations are independent.
3. The form of the model is correct.
 - ▶ Linear relationship between logits and explanatory variables
 - ▶ All relevant variables are included; irrelevant ones excluded
4. Sample size is large enough for valid inference-tests and CIs.
(Recall large-sample properties of MLEs.)

Var is not constant

Binary Logistic Regression vs Linear Regression

- ▶ Both utilize MLE's for the β 's
- ▶ *Less assumptions to check for than in linear (least squares) regression*
 - ▶ No need to check for outliers since Y is either 0 or 1.
 - ▶ No residual plots; No meaning can be inferred from residuals
 - ▶ Variance is not constant

Case Study III: Testing model assumptions

► Independence: We know that there were families within Donner's party, so we have concerns that the observations were not independent!

⚠ Form of the model: Test higher-order terms such as

- ▶ Age^2 - non-linear (quadratic) in X
- ▶ Sex * Age interaction, and
- ▶ $\text{Age}^2 * \text{Sex}$ interaction.

Comparing models: Likelihood Ratio Test

- Idea: Compare likelihood of data under FULL (F) model, \mathcal{L}_F to likelihood under REDUCED (R) model, \mathcal{L}_R of same data.

Likelihood ratio : $\frac{\mathcal{L}_R}{\mathcal{L}_F}$, where $\mathcal{L}_R \leq \mathcal{L}_F$

- Hypotheses: $H_0 : \beta_1 = \cdots = \beta_k = 0$

(Reduced model is appropriate; fits data as well as Full model)

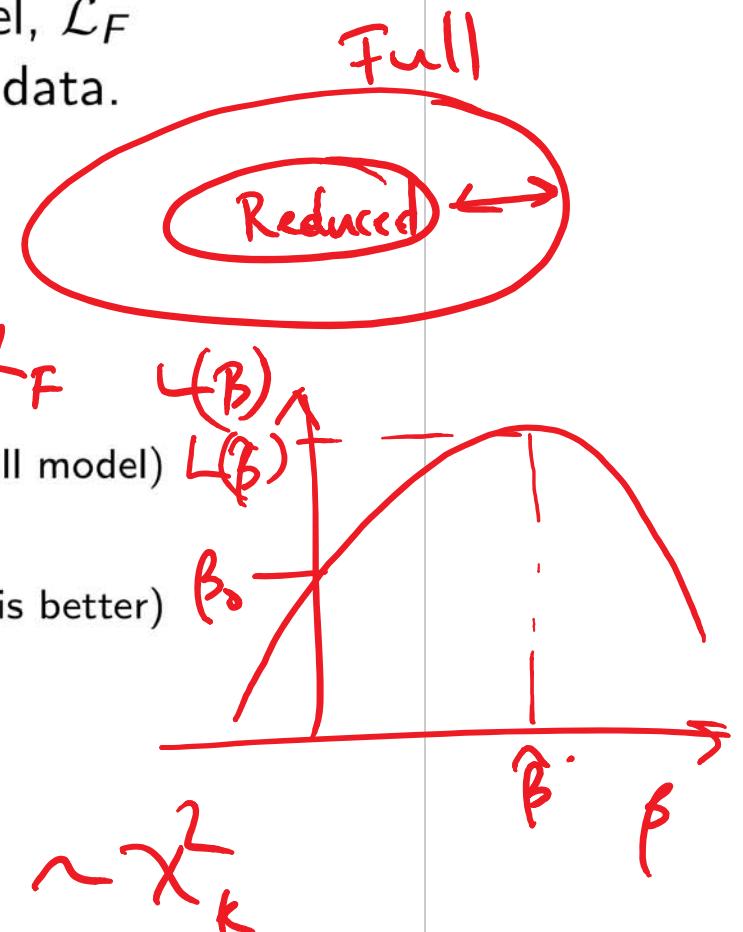
H_a : at least one $\beta_1, \dots, \beta_k \neq 0$

(Full model is better)

- Test Statistic: Deviance (residual),

$$G^2 = [-2 \log \mathcal{L}_R] - [-2 \log \mathcal{L}_F] = -2 \log \left(\frac{\mathcal{L}_R}{\mathcal{L}_F} \right) \sim \chi^2_k$$

- For large n , under H_0 , G^2 is an observation from a chi-square distribution with k df.



Case Study III Exercise: Comparing models

Using R output,

Q: Determine whether a model with the 3 higher-order polynomial terms and/or interaction terms is an improvement over the additive model.

- ▶ Hypotheses:
- ▶ Test Statistic:
- ▶ Distribution of TS:
- ▶ P-value:
- ▶ Conclusion:

Case Study 3: Higher Order Model with 3 higher order/interaction terms

```
fitfull<-glm(Status~Age+sex+Age:sex+I(Age^2)+I(Age^2):sex, family=binomial, data=donner)
summary(fitfull)

##
## Call:
## glm(formula = Status ~ Age + sex + Age:sex + I(Age^2) + I(Age^2):sex,
##      family = binomial, data = donner)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.3396 -0.9757 -0.3438  0.5269  1.5901
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.318484   3.940184 -0.842   0.400
## Age                      0.183031   0.226632  0.808   0.419
## sexFemale                0.265286  10.455222  0.025   0.980
## I(Age^2)                 -0.002803   0.002985 -0.939   0.348
## Age:sexFemale             0.299877   0.696050  0.431   0.667
## sexFemale:I(Age^2)        -0.007356   0.010689 -0.688   0.491
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 45.361 on 39 degrees of freedom
## AIC: 57.361
```

Testing β 's: Wald versus LRT test

	Wald	LRT
Testing whether a single $\beta=0$		
Comparing nested models		
Small to moderate sample sizes β near boundary of parameter space		

Binary Logistic Regression

Case Study III Exercise: Comparing models

Using R output,

Q: Determine whether the effect of Age on the odds of survival differ with Sex.

- Hypotheses:

$$H_0: \text{logit}(\pi) = \beta_0 + \beta_1 \text{Age} + \beta_2 I_F \quad \Rightarrow \quad \beta_3 = 0$$

$$H_a: \text{logit}(\pi) = \beta_0 + \beta_1 \text{Age} + \beta_2 I_F + \beta_3 (\text{Age} \times I_F)$$

- Test Statistic:

$$\chi^2 = D_0 - D_A = 51.256 - 47.346 = 3.91 |$$

- Distribution of TS:

$$\chi^2 \approx 3.91 \sim \chi^2_1$$

- P-value:

$$P(\chi^2_1 > 3.91) = 0.048$$

- Conclusion:

Suggestive but not conclusive evidence of interaction

Case Study 3: Interaction Model, Age*Sex

```
fitas<-glm(Status~Age*sex, family=binomial, data=donner)
summary(fitas)

##
## Call:
## glm(formula = Status ~ Age * sex, family = binomial, data = donner)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.2279  -0.9388  -0.5550   0.7794   1.6998
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.31834   1.13103   0.281   0.7784
## Age         -0.03248   0.03527  -0.921   0.3571
## sexFemale    6.92805   3.39887   2.038   0.0415 *
## Age:sexFemale -0.16160   0.09426  -1.714   0.0865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Dispersion parameter for binomial family taken to be 1
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 47.346 on 41 degrees of freedom
## AIC: 55.346
##
```

Comparing models: ‘Global’ LRT

Case Study III Exercise: 'Global' LRT

Using R output,

Q: Determine whether or not the additive model fits better than the Null model.

- ▶ Hypotheses:
- ▶ Test Statistic:
- ▶ Distribution of TS:
- ▶ P-value:
- ▶ Conclusion:

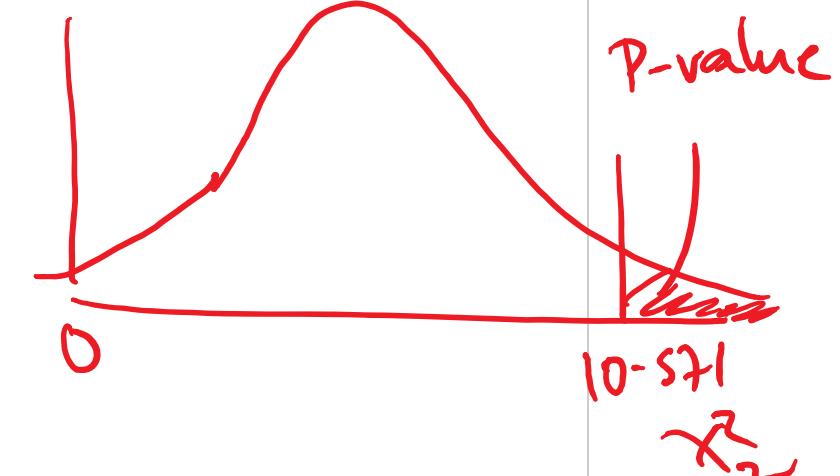
$$\text{logit}(\pi) = \beta_0$$

$$H_0: (\text{Null}) \quad \beta_1 = \beta_2 = 0 \quad \text{vs} \quad H_A: (\text{additive model is better})$$

$$Q^2 = D_N - D_A = 61.827 - 51.256 = 10.571 \sim \chi^2_2$$

at least 1 of the β 's not zero

The additive model is better. IOW, Age and Sex are relevant for the log odds of survival.



Case Study 3: Additive model for Survived

```
fitasf<-glm(Status~Age+sex, family=binomial, data=donner)
summary(fitasf)
```

```
##
## Call:
## glm(formula = Status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q      Max
## -1.7445 -1.0441 -0.3029  0.8877  2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471  0.1413
## Age        -0.07820   0.03728  -2.097  0.0359 *
## sexFemale   1.59729   0.75547   2.114  0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```



$$H_0: \beta_j = 0$$

$$H_a: \beta_j \neq 0$$

$$\hat{\beta}_j$$

$$SE(\hat{\beta}_j)$$

$$\hat{\beta}_j / SE(\hat{\beta}_j)$$

$$2P(Z \geq |z|)$$

$$\hat{\beta}_1$$

$$\hat{\beta}_2$$

$$R : \text{Deviance}_R = -2 \ln L_R$$

$$F : \text{Deviance}_F = -2 \ln L_F$$

$$G^2 = \text{Deviance}_R - \text{Deviance}_F$$

$$= 61.827 - 51.256$$

Week 4 R functions

- ▶ Create factor: `as.factor()`
- ▶ Cross Tabulations: `xtabs()`
- ▶ Specifying the reference level: `relevel()`
- ▶ Generalized Linear Models: `glm()`
- ▶ Find deviance: `deviance()`

STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



February 6-8, 2018

1/34

STA 303/1002: Class 10- Logistic Regression

- ▶ Case Study III: The Donner Party Example
 - ▶ Comparing models
 - ▶ Wald vs Likelihood Ratio Tests
 - ▶ Effect Plots
 - ▶ Related R packages and functions
- ▶ Case Study IV: Binomial Logistic Regression

Logistic Regression: Testing whether single β 's are zero

WALD CHI-SQUARE PROCEDURES

- ▶ **Hypotheses:** $H_0 : \beta_j = 0$ (X_j has no effect on log-odds)
 $H_a : \beta_j \neq 0$

- ▶ **Test Statistic:**
$$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

where

- ▶ $\hat{\beta}_j$ - maximum likelihood (ML) estimate and
- ▶ $SE(\hat{\beta}_j)$ - estimated standard error from the numerical procedure that generated the MLE.
- ▶ By standard large-sample results, MLE's are normally distributed. Thus, for large n , under H_0 , z is an observation from an approx. $\mathcal{N}(0, 1)$ distribution.
- ▶ **95% Confidence interval:**
$$\hat{\beta}_j \pm 1.96SE(\hat{\beta}_j)$$

Examples: Testing whether β 's are zero

Using R output:

	Age	Sex
Test statistic	$(-0.078/0.0373)^2$	4.47
P-value	0.036	0.0345
95% CI for β	$-0.078 \pm 1.96(0.0373)$ $=(-0.15, -0.0055)$	(0.117, 3.078)
CI for Odds ratio	$(e^{-0.15}, e^{-0.0055}) = (0.86, 0.995)$	(1.124, 21.72)
Conclusion	For the same sex, the odds ratio for a 1-year increase in age is between .86 and 0.995.	

- ▶ Note: Both marginal p-values are less than 0.05 and the confidence intervals for the odds ratios do not include 1.
- ▶ Hence, we have moderate evidence that both *Age* and *Sex* have an effect on survival over and above each other.
- ▶ Recall: If $Z \sim \mathcal{N}(0, 1)$, then $Z^2 \sim \chi_1$.

Model Assumptions for Binary Logistic Regression

1. Underlying probability model for response is Bernoulli.
2. Observations are independent.
3. The form of the model is correct.
 - ▶ Linear relationship between logits and explanatory variables
 - ▶ All relevant variables are included; irrelevant ones excluded
4. Sample size is large enough for valid inference-tests and CIs.
(Recall large-sample properties of MLEs.)

Comparing models: Likelihood Ratio Test

- ▶ Idea: Compare likelihood of data under FULL (F) model, \mathcal{L}_F to likelihood under REDUCED (R) model, \mathcal{L}_R of same data.

Likelihood ratio : $\frac{\mathcal{L}_R}{\mathcal{L}_F}$, where $\mathcal{L}_R \leq \mathcal{L}_F$

- ▶ Hypotheses: $H_0 : \beta_1 = \cdots = \beta_k = 0$ \geq 1 \beta.

(Reduced model is appropriate; fits data as well as Full model)

- H_a : at least one $\beta_1, \dots, \beta_k \neq 0$

(Full model is better)

- ▶ Test Statistic: Deviance (residual),

$$G^2 = -2 \log \mathcal{L}_R - (-2 \log \mathcal{L}_F) = -2 \log \left(\frac{\mathcal{L}_R}{\mathcal{L}_F} \right)$$

$$G^2 = D_R - D_F$$

\uparrow

- ▶ For large n , under H_0 , G^2 is an observation from a chi-square distribution with \underline{k} df.

6/34

Case Study III Exercise: Comparing models

Using R output,

Q: Determine whether a model with the 3 higher-order polynomial terms and/or interaction terms is an improvement over the additive model.

H_0 : Additive vs H_a : Higher order model

- Hypotheses:

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0$$

- Test Statistic:

$$G^2 = 51.256 - 45.361 = 5.895 \sim \chi^2_3$$

- Distribution of TS:

- P-value:

$$P(\chi^2_3 > 5.895) = 0.1168$$

- Conclusion:

Additive model is satisfactory.

Case Study 3: Higher Order Model with 3 higher order/interaction terms

```
fitfull<-glm(Status~Age+sex+Age:sex+I(Age^2)+I(Age^2):sex, family=binomial, data=donner)
summary(fitfull)
```

```
##  
## Call:  
## glm(formula = Status ~ Age + sex + Age:sex + I(Age^2) + I(Age^2):sex,  
##       family = binomial, data = donner)  
##  
## Deviance Residuals:  
##      Min        1Q    Median        3Q       Max  
## -2.3396  -0.9757  -0.3438   0.5269   1.5901  
##  
## Coefficients:  
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -3.318484  3.940184 -0.842   0.400  
## Age          0.183031  0.226632  0.808   0.419  
## sexFemale    0.265286 10.455222  0.025   0.980  
## I(Age^2)     -0.002803  0.002985 -0.939   0.348  
## Age:sexFemale 0.299877  0.696050  0.431   0.667  
## sexFemale:I(Age^2) -0.007356  0.010689 -0.688   0.491  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 61.827 on 44 degrees of freedom  
## Residual deviance: 45.361 on 39 degrees of freedom  
## AIC: 57.361
```

$$D_F = 45.361$$

Testing β 's: Wald versus LRT test

	Wald	LRT	
Testing whether a single $\beta=0$	✓ (easier)	✓	(The two methods are different.)
Comparing nested models ($>1 \beta$)		✓	
Small to moderate sample sizes β near boundary of parameter space		✓	(More reliable in general)

related to
normality and
regularity conditions
of MLEs.

Case Study III Exercise: Comparing models

Using R output,

Q: Determine whether the effect of Age on the odds of survival differ with Sex.

$$H_0: (\text{Additive}) \quad \text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 \mathbf{1}_F \Leftrightarrow H_0: \gamma_3 = 0$$

► Hypotheses:

$$H_a: (\text{Interaction}) \quad \text{logit}(\hat{\pi}) = \hat{\gamma}_0 + \hat{\gamma}_1 \text{Age} + \hat{\gamma}_2 \mathbf{1}_F + \hat{\gamma}_3 \text{Age} * \mathbf{1}_F$$

► Test Statistic:

$$\underline{\text{LRT}}$$

$$\chi^2 = 51.256 - 47.346$$

► Distribution of TS: $= 3.91 \sim \chi^2_1$

► P-value: $P\text{-value} = P(\chi^2_1 > 3.91)$

► Conclusion: $= 0.048$ (using R)

$$\underline{\text{Wald}}$$

$$z = \hat{\gamma}_3 / \hat{s.e}(\hat{\gamma}_3)$$

$$= -1.714 \sim N(0,1).$$

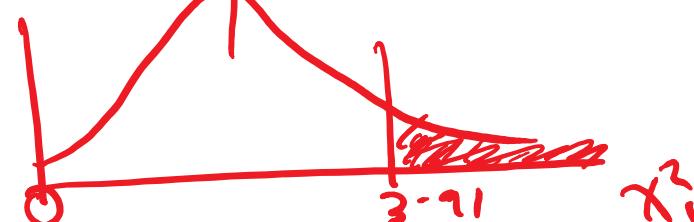
$$P\text{-value} = 0.0865$$

Suggestive but inconclusive evidence of interaction.

Binary Logistic Regression

$$1 - \text{pchisq}(3.91, df=1)$$

10/34

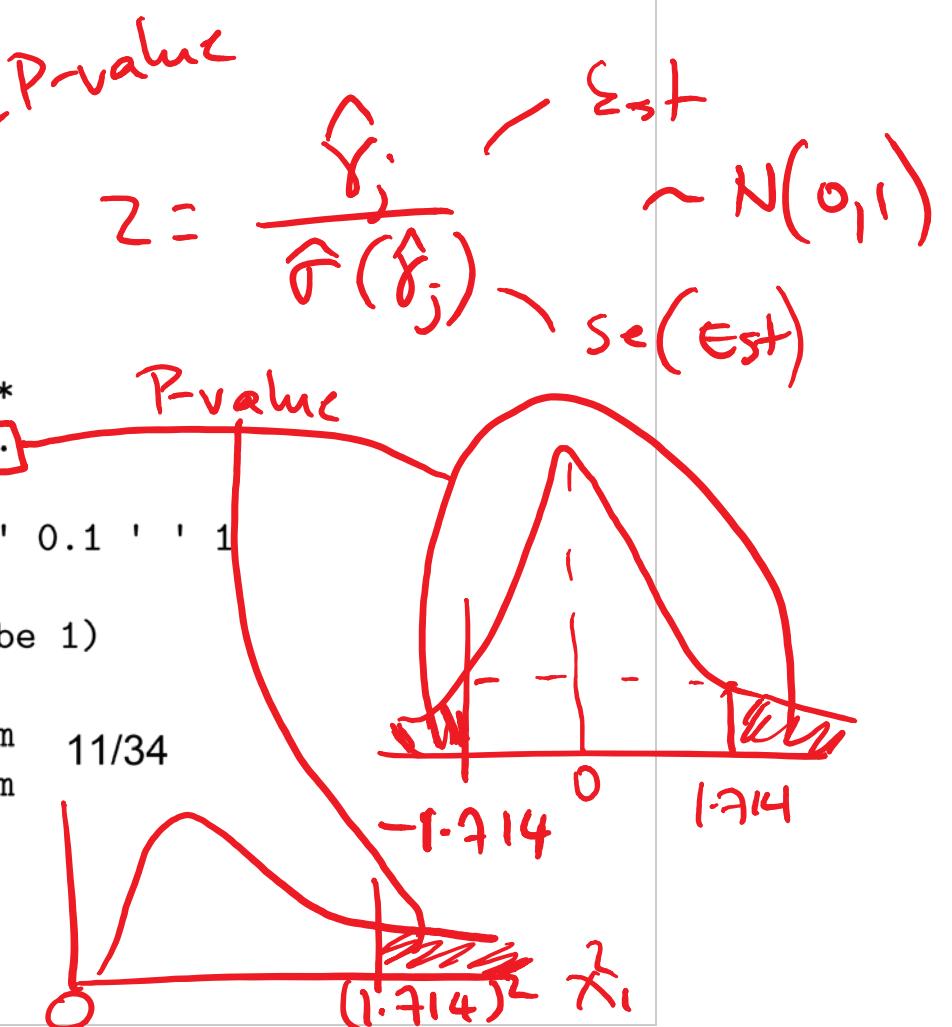


Case Study 3: Interaction Model, Age*Sex

```
fitas<-glm(Status~Age*sex, family=binomial, data=donner)
summary(fitas)
```

```
##
## Call:
## glm(formula = Status ~ Age * sex, family = binomial, data = donner)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.2279  -0.9388  -0.5550   0.7794   1.6998
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.31834   1.13103   0.281   0.7784
## Age         -0.03248   0.03527  -0.921   0.3571
## sexFemale   6.92805   3.39887   2.038   0.0415 *
## Age:sexFemale -0.16160   0.09426  -1.714   0.0865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Dispersion parameter for binomial family taken to be 1
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 47.346 on 41 degrees of freedom
## AIC: 55.346
```

Note: Z-procedure & χ^2 procedure are equivalent



Comparing models: ‘Global’ LRT

- ▶ **Idea:** Compares Fitted model to NULL [$\text{logit}(\pi) = \beta_0$] model
- ▶ **Hypotheses:** $H_0 : \beta_1 = \cdots = \beta_p = 0$
(NULL model is appropriate)
 $H_a : \text{at least one } \beta_1, \dots, \beta_p \neq 0$
(Fitted model is better)

$$\hat{\chi}^2 = D_N - D_{\text{Fitted}} \sim \chi^2_p$$

12/34

Case Study III Exercise: 'Global' LRT

Using R output,

Q: Determine whether or not the additive model fits better than the Null model.

- ▶ Hypotheses:
- ▶ Test Statistic:
- ▶ Distribution of TS:
- ▶ P-value:
- ▶ Conclusion:

Done.
(See previous class)
($P=0.005$)

Case Study 3: Additive model for Survived

```
fitasf<-glm(Status~Age+sex, family=binomial, data=donner)
summary(fitasf)

##
## Call:
## glm(formula = Status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.63312   1.11018   1.471   0.1413
## Age         -0.07820   0.03728  -2.097   0.0359 *
## sexFemale   1.59729   0.75547   2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 61.827 on 44 degrees of freedom
## Residual deviance: 51.256 on 42 degrees of freedom  14/34
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

$$D_R = 51.256$$

STA303/1004 - Class 10 R Markdown

February 6, 2018

Case 3: Deviance test and Estimated Var-Cov of β

```
anova(fitasf, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL          44    61.827
## Age     1   5.5358    43  56.291  0.01863 *
## sex     1   5.0344    42  51.256  0.02485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(vcov(fitasf,digits=3))
```



	(Intercept)	Age	sexFemale
(Intercept)	1.23250837	-0.038472741	0.06007099
Age	-0.03847274	0.001390134	-0.00823197
sexFemale	0.06007099	-0.008231970	0.57073339

Var-Cov matrix for $\hat{\beta}_j$'s.

16/34

$$\text{Eq, } \sqrt{0.57} = 0.757.$$

$$\hat{\sigma}^2(\hat{\beta}_j) = (\hat{se}(\hat{\beta}_j))^2$$

Case 3: Confidence Intervals for β 's

```
cbind(bhat=coef(fitasf), confint.default(fitasf)) # 95% CI for betas
```

```
##          bhat    2.5 %    97.5 %
## (Intercept) 1.63312031 -0.5428002 3.809040837
## Age         -0.07820407 (-0.1512803 -0.005127799)
## sexFemale   1.59729350 (0.1166015 3.077985503)
```

$$\hat{\beta}_j \pm 1.96 \text{ se}(\hat{\beta}_j)$$

Compare to results
on p. 4

```
exp(coef(fitasf)) # exponentiate estimated betas, get odds ratios
```

```
## (Intercept)      Age     sexFemale
## 5.1198252     0.9247757 4.9396452
```

```
exp(cbind(OR=coef(fitasf), confint.default(fitasf))) #CI for odds ratio
```

```
##          OR    2.5 %    97.5 %
## (Intercept) 5.1198252 0.5811187 45.1071530
## Age         0.9247757 (0.8596067 0.9948853)
## sexFemale   4.9396452 (1.1236716 21.7146143)
```

Case 3: Wald tests in R

Computes Wald chi-squared test for 1 or more β coefficients

- ▶ R package: aod (Analysis of Overdispersed Data)
- ▶ Syntax `wald.test(Sigma, b, Terms)`
- ▶ Sigma: var-cov matrix, extracted from the `glm` function
- ▶ b: coefficients (`coef(glm())`)
- ▶ Terms: specifies which terms in the models are to be tested

```
library(aod) # Analysis of Overdispersed Data  
wald.test(Sigma=vcov(fitasf), b=coef(fitasf), Terms=2:3)
```

```
## Wald test:  
## -----  
##  
## Chi-squared test:  
## X2 = 6.9, df = 2, P(> X2) = 0.032
```

$\hat{\beta}_j$

$\beta_1 \text{Age}$
 $\beta_2 \text{I}_{F \cdot}$

Compare to $\chi^2 \cdot 13$ with LRT
($P = 0.005$)

Case 3: Wald tests in R

```
# Testing interaction, Refer to interaction model  
# summary(fitas)  
# Testing a single beta  
wald.test(Sigma=vcov(fitas), b=coef(fitas), Terms=4)
```

```
## Wald test:  
## -----  
##  
## Chi-squared test:  
## X2 = 2.9, df = 1, P(> X2) = 0.086
```

$$H_0: \gamma_3 = 0$$
$$H_a: \gamma_3 \neq 0$$

Compare with
 $P > 0.1$

Case 3: Estimated probability of survival

$$y_i = 0, 1 \quad (\text{observed})$$

$$\begin{aligned} \text{logit } (\pi) &= \beta_0 + \beta_1 \text{Age} + \beta_2 I_F \\ \ln \left(\frac{\hat{\pi}}{1-\hat{\pi}} \right) &= \hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 I_F \quad \Rightarrow \hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 I_F}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 I_F}} \end{aligned}$$

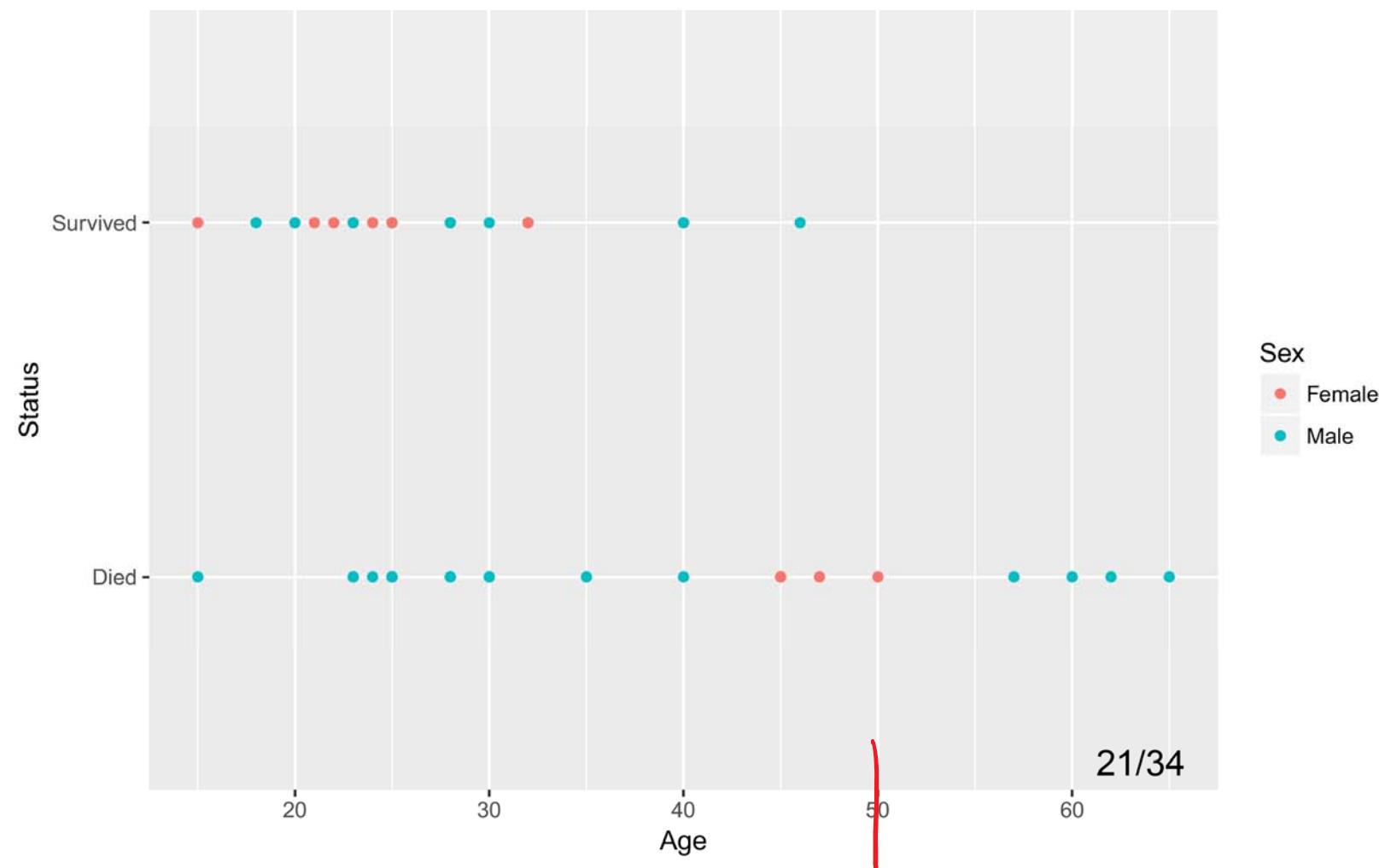
phats<-predict.glm(fitasf, type="response") # predicted probability of survival
phats[1:5] ↑ ↑ (Estimate).

```
##      1      2      3      4      5  
## 0.4587010 0.5255405 0.1831661 0.3289359 0.3643360
```

$$0 < \pi < 1$$

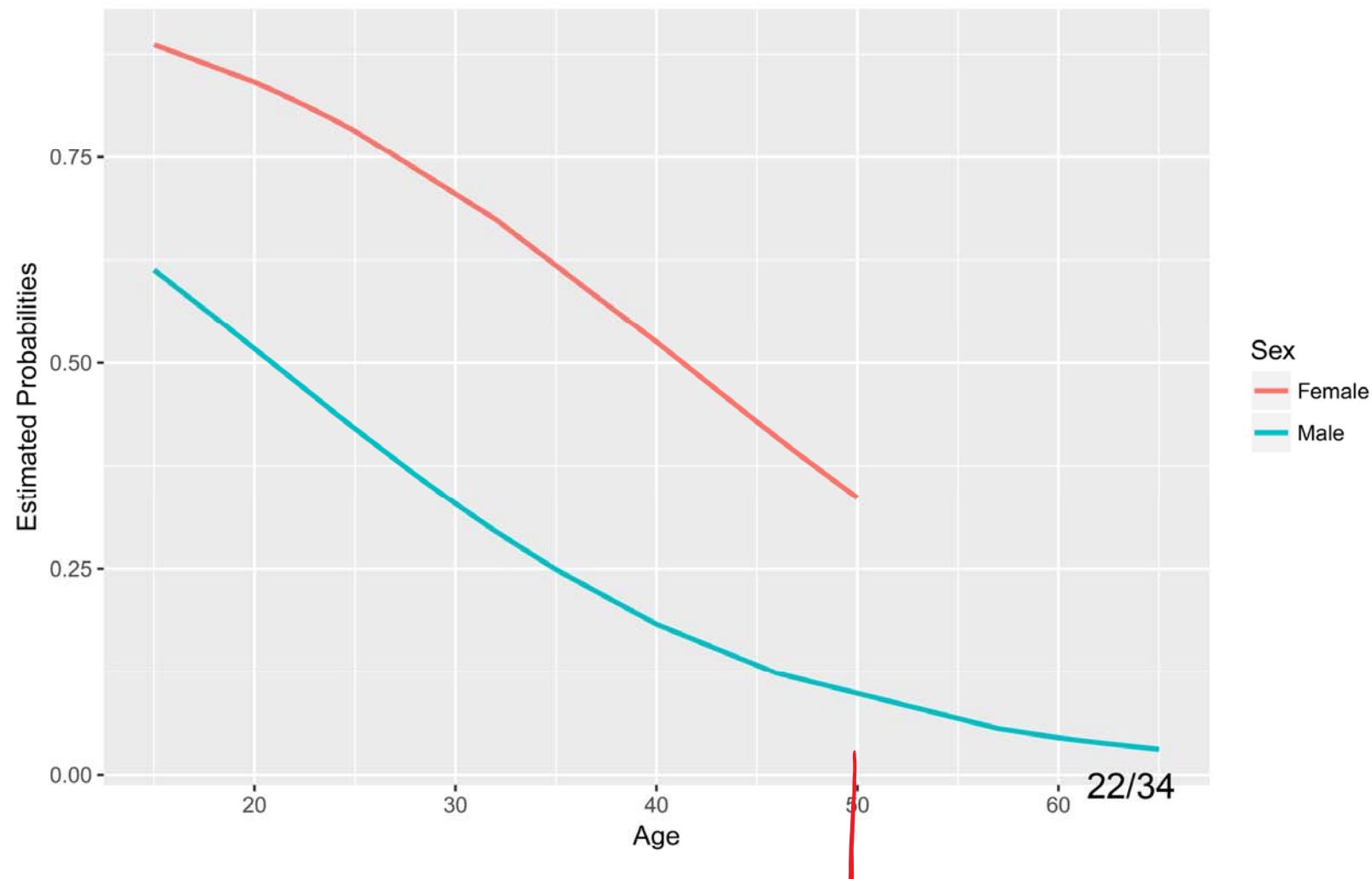
Case 3 Plots: Of Data

```
#contrasts(Status)
library(ggplot2)
ggplot(donner, aes(x=Age, y>Status, color=Sex))+geom_point()
```



Case 3 Plots: Additive Logistic Regression Model

```
ggplot(donne, aes(x=Age, y=phats)) + ylab("Estimated Probabilities") +  
  geom_line(aes(color=Sex), size=1)
```

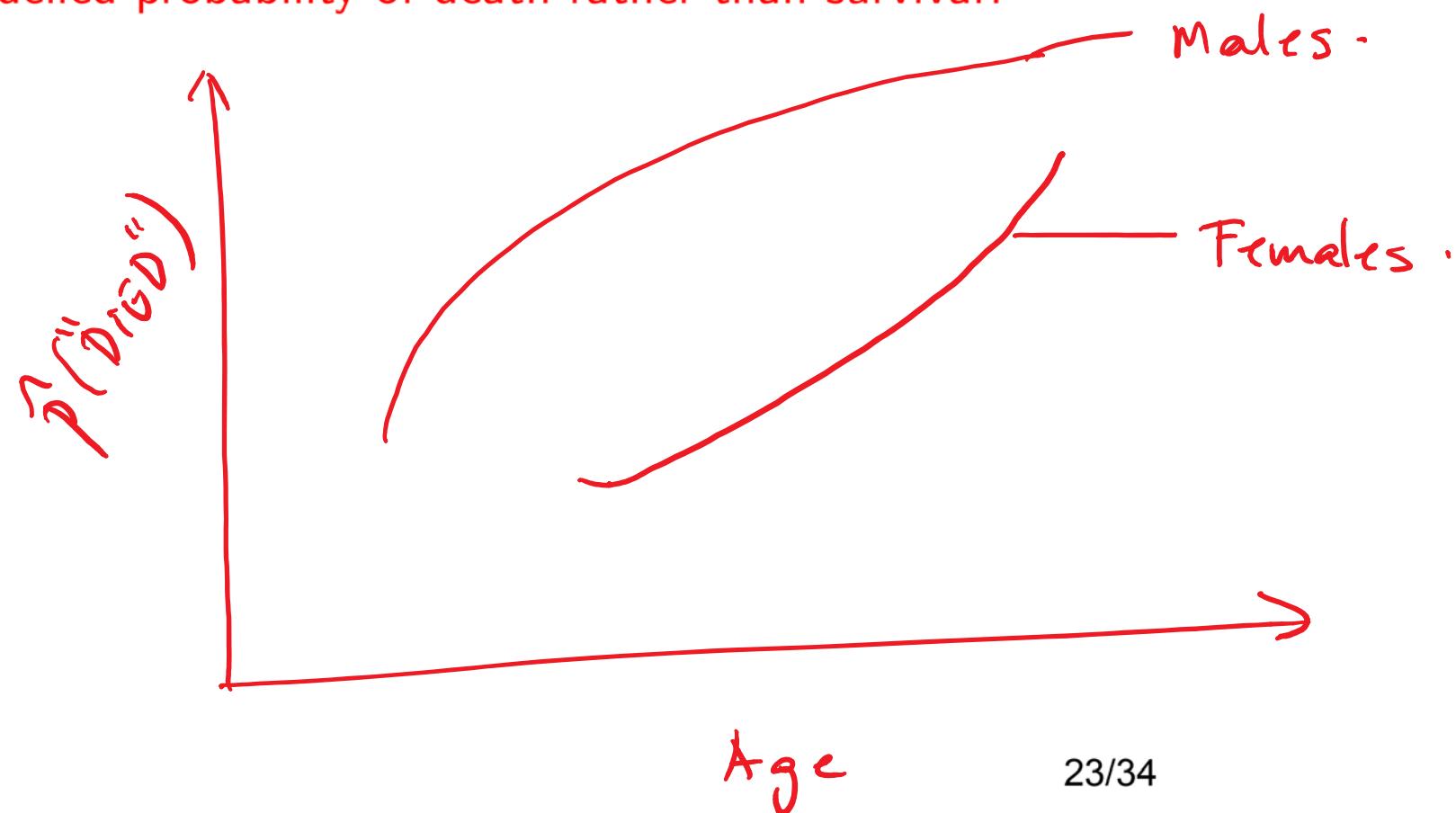


Plot

$$\hat{P}(\text{"SURVIVED"}) = 1 - \hat{P}(\text{"DIED"})$$

Q: How would the plot of estimated probabilities change if we modelled probability of death rather than survival?

$$\text{ODDS(SURVIVED)} = \frac{1}{\text{ODDS(DIED)}}$$



Logistic Regression

Over 50yrs

Q: Should one be reluctant to draw conclusions about the ratio of male and female odds of survival for the Donner Party members over 50?

Yes, since there were no women older than 50.
the model may not extend.

Other Model Fit Statistics

- ▶ Two popular fit statistics: AIC and BIC; combines log-likelihood with a penalty.
- ▶ Useful for comparing models with same response and same data
- ▶ Extends from normal regression to GLMs
 1. Akaike's Information Criterion (AIC)

$$AIC = -2 \log \mathcal{L} + 2(p + 1)$$

2. Schwarz's (Bayesian Information) Criterion (BIC)

$$BIC = -2 \log \mathcal{L} + (p + 1) \log N$$

where

- ▶ p =number of explanatory variables, and
- ▶ N =sample size

25/34

Model Fit Statistics: AIC and BIC

- ▶ Smaller is better!
- ▶ BIC applies stronger penalty for model complexity than AIC
- ▶ AIC Rule of Thumb:
 - ▶ One model fits **better** than another if difference in AIC's > 10
 - ▶ One model is essentially **equivalent** to another if the difference in AIC's < 2

Using AIC: Case Study III Example

- ▶ Fitted models are based on same response and data.
- ▶ Based on AIC, choose a ‘best’ model.

Model	Variables	AIC	BIC
1	{age,sex}	57.256	62.676
2	{age,sex,age*sex,age ² ,age ² *sex}	57.361	68.201
3	{age,sex,age*sex,age ² }	55.830	64.863
4	{age,sex,age*sex}	55.346	62.573

BIC (fitted model)

Results:

- ▶ Difference in AIC between 1 and 3 is within 2
- ▶ There is some indication that 2 is worse than 3 and 4.
- ▶ Choose Model 1 (the simplest)

27/34

Related R packages and functions

- ▶ **Packages:**

- ▶ aod: analysis of over-dispersed data
- ▶ ggplot2: graphics
- ▶ Sleuth3: data sets for Ramsey and Schafer's text
- ▶ effects: effects displays for GLM and other models

- ▶ **Functions:**

- ▶ confint()
- ▶ coef()
- ▶ vcov()
- ▶ wald.test()
- ▶ AIC()
- ▶ BIC()

Binomial Logistic Regression

29/34

Suppose $Y \sim \text{Binomial}(m, \pi)$

- ▶ Y -binomial count of the number of “successes”

$$P(Y = y) = \binom{m}{y} \pi^y (1 - \pi)^{m-y}, \quad y = 0, 1, \dots, \underline{m}$$

- ▶ Link to Bernoulli:
 $Y = \sum_{i=1}^m X_i$ if X_i 's are independent $\text{Bernoulli}(\pi)$ r.v.s.
Assume that π is the same for each Bernoulli trial.
- ▶ Mean: $E(Y) = m\pi$
- ▶ Variance: $\text{Var}(Y) = m\pi(1 - \pi)$

Suppose $Y \sim \text{Binomial}(m, \pi)$

- ▶ Consider modelling

$$P = \frac{Y}{m}$$

- the proportion of “successes” out of m independent Bernoulli trials.

- ▶ where,

- ▶ $E\left(\frac{Y}{m}\right) = \pi$

- ▶ $\text{Var}\left(\frac{Y}{m}\right) = \frac{\pi(1 - \pi)}{m}$

Case Study IV Data Example

- ▶ Data: counts of bird species for 18 Krunnit Islands off Finland.

i =	x_i area	m_i nspecies	y_i nextinct
ISLAND	AREA	ATRISK	EXTINCT
Ulkokrunni	185.8	75	5
Maakrunni	105.8	67	3
Ristikari	30.7	66	10
Isonkivenletto	8.5	51	6
...			
Tiirakari	0.2	40	13
Ristikarenletto	0.07	6	3

$$p_c = \frac{y_i}{m_i}$$

- ▶ AREA- area of island in km^2 , x_i
- ▶ ATRISK- number of species on each island in 1949, m_i
- ▶ EXTINCT- number of species no longer found on each island in 1959, y_i

Case Study IV: Model

- ▶ π_i - probability of 'extinction' of each island.
Assume that this is the same for each species of bird on a particular island.
- ▶ *Assume species survival is independent.* Then

$$Y_i \sim \text{Binomial}(m_i, \pi_i)$$

- ▶ Unlike Case III- Donner party binary logistic example, we can estimate π_i from the data.

Case Study IV: Model

- ▶ Observed response proportion:

$$\bar{\pi}_i = \frac{y_i}{m_i}$$

- ▶ Observed or Empirical logits: (S-“saturated”)

$$\log\left(\frac{\bar{\pi}_{S,i}}{1 - \bar{\pi}_{S,i}}\right) = \log\left(\frac{y_i}{m_i - y_i}\right)$$

- ▶ Proposed Model: $\boxed{\log\left(\frac{\pi_{S,i}}{1 - \pi_{S,i}}\right) = \beta_0 + \beta_1 Area_i,} \quad i = 1, \dots, 18$

- ▶ AIM:

- ▶ Learn how to create nature preserves that help endangered species.
- ▶ Are large or small preserves better?

34/34

STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



February 6-8, 2018

1/32

STA 303/1002: Class 11- Binomial Logistic Regression

- ▶ Case Study IV: Island size and bird extinction
 - ▶ R syntax
 - ▶ Data visualization
 - ▶ Interpreting coefficients
 - ▶ Wald procedures
- ▶ Principle of the week: *K-Keep, I-It, S-Simple, S-Stupid*(US Navy, 1960)



2/32

Plot

Q: How would the plot of estimated probabilities change if we modelled probability of death rather than survival?

3/32

Over 50yrs

Q: Should one be reluctant to draw conclusions about the ratio of male and female odds of survival for the Donner Party members over 50?

Other Model Fit Statistics

- ▶ Two popular fit statistics: AIC and BIC; combines log-likelihood with a penalty.
- ▶ Useful for comparing models with same response and same data
- ▶ Extends from normal regression to GLMs
 1. Akaike's Information Criterion (AIC)

$$AIC = -2 \log \mathcal{L} + 2(p + 1)$$

2. Schwarz's (Bayesian Information) Criterion (BIC)

$$BIC = -2 \log \mathcal{L} + (p + 1) \log N$$

where

- ▶ p -number of explanatory variables, and
- ▶ N =sample size

Model Fit Statistics: AIC and BIC

- ▶ Smaller is better!
- ▶ BIC applies stronger penalty for model complexity than AIC
- ▶ AIC Rule of Thumb:
 - ▶ One model fits **better** than another if difference in AIC's > 10
 - ▶ One model is essentially **equivalent** to another if the difference in AIC's < 2

Using AIC: Case Study III Example

- ▶ Fitted models are based on same response and data.
- ▶ Based on AIC, choose a ‘best’ model.

Model	Variables	AIC	BIC
1	{age,sex}	57.256	62.676
2	{age,sex,age*sex,age ² ,age ² *sex}	57.361	68.201
3	{age,sex,age*sex,age ² }	55.830	64.863
4	{age,sex,age*sex}	55.346	62.573

Results:

- ▶ Difference in AIC between 1 and 3 is within 2
- ▶ There is some indication that 2 is worse than 3 and 4.
- ▶ Choose Model 1 (the simplest)

7/32

Related R packages and functions

- ▶ **Packages:**

- ▶ aod: analysis of over-dispersed data
- ▶ ggplot2: graphics
- ▶ Sleuth3: data sets for Ramsey and Schafer's text
- ▶ effects: effects displays for GLM and other models

- ▶ **Functions:**

- ▶ confint()
- ▶ coef()
- ▶ vcov()
- ▶ wald.test()
- ▶ AIC()
- ▶ BIC()

Binomial Logistic Regression

9/32

Suppose $Y \sim \text{Binomial}(m, \pi)$

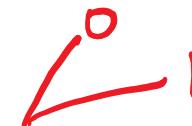
- ▶ Y -binomial count of the number of “successes”

$$P(Y = y) = \binom{m}{y} \pi^y (1 - \pi)^{m-y}, \quad y = 0, 1, \dots, m$$

- ▶ Link to Bernoulli:

$Y = \sum_{i=1}^m X_i$ if X_i 's are independent Bernoulli(π) r.v.s.

Assume that π is the same for each Bernoulli trial.



- ▶ Mean: $E(Y) = m\pi$

- ▶ Variance: $\text{Var}(Y) = m\pi(1 - \pi)$

Suppose $Y \sim \text{Binomial}(m, \pi)$

- ▶ Consider modelling

$$\frac{Y}{m}$$

- the proportion of “successes” out of m independent Bernoulli trials.

- ▶ where,

- ▶ $E\left(\frac{Y}{m}\right) = \pi$

- ▶ $\text{Var}\left(\frac{Y}{m}\right) = \frac{\pi(1 - \pi)}{m}$

Case Study IV Data Example

- ▶ Data: counts of bird species for 18 Krunnit Islands off Finland.

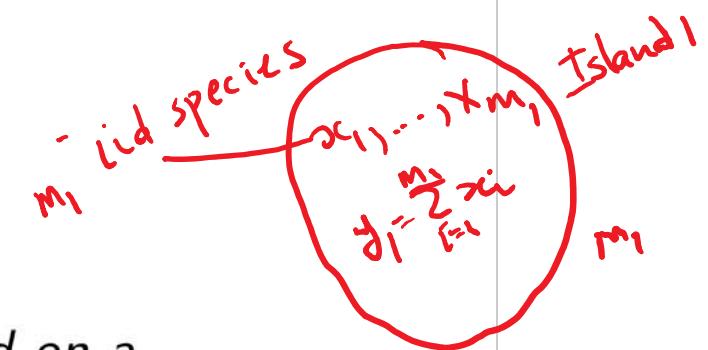
i	x_i area	m_i nspecies	y_i nextinct	# of successes \bar{x}_i	observed proportion s/m
ISLAND	AREA	ATRISK	EXTINCT		
Ulkokrunni	185.8	75	5		
Maakrunni	105.8	67	3		
Ristikari	30.7	66	10		
Isonkivenletto	8.5	51	6		
...					
Tiirakari	0.2	40	13		
Ristikarenletto	0.07	6	3		

\bar{x}_i
 $5/75$
 $3/67$
 \vdots
 $13/40$
 $3/6 = 0.5$

- ▶ AREA- area of island in km^2 , x_i
- ▶ ATRISK- number of species on each island in 1949, m_i
- ▶ EXTINCT- number of species no longer found on each island in 1959, y_i

Case Study IV: Model

- ▶ π_i - probability of 'extinction' for each island.
- ① Assume that this is the same for each species of bird on a particular island.
- ② ▶ Assume species survival is independent. Then



$$0 \leq \pi_i \leq 1$$

$$0 < \pi_i < 1$$

- ▶ Unlike Case III- Donner party binary logistic example, we can estimate π_i from the data.

Bernoulli
 → Binomial counts
 → $\pi = \{0, 1\}$
 Proportions vs Percentages
 (0, 1)
 (0, 100%)
 cts .

Case Study IV: Model

Data

- ▶ Observed response proportion:

$$\bar{\pi}_i = \frac{y_i}{m_i} \quad \begin{matrix} \text{observed counts} \\ \text{total} \end{matrix} \rightarrow \bar{\pi}_i$$

- ▶ Observed or Empirical logits: (S-“saturated”)

$$\log\left(\frac{\bar{\pi}_{S,i}}{1 - \bar{\pi}_{S,i}}\right) = \log\left(\frac{y_i}{m_i - y_i}\right)$$

Estimates

- ▶ Proposed Model: $\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 Area_i, \quad i = 1, \dots, 18$

$$\hat{\pi}_i$$

- ▶ AIM:

- ▶ Learn how to create nature preserves that help endangered species.
- ▶ Are large or small preserves better?

Case Study IV: Initial assessment of data

Visuals

- ▶ Plot observed logits versus area to see if a linear relationship seems appropriate.
- ▶ From that plot, we decide to look at log(Area) instead.
- ▶ The relationship between empirical logits and log(Area) seems linear.
- ▶ Hence, we fit

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \log(Area_i), \quad i = 1, \dots, 18$$

Case Study IV: R syntax

- ▶ In R, the model formula has the form:

$$\text{cbind}(y_i, m_i - y_i) \sim \text{log(Area)}$$

Need to specify both:

- ▶ y_i - number of successes and
- ▶ $(m_i - y_i)$ - number of failures



Case Study IV: Model Summary

- ▶ Number of observations: 18
- ▶ Number of coefficients: 2
- ▶ Fitted model:

$$\text{logit } (\hat{\pi}) = -1.196 - 0.297 \log(\text{Area})$$

$\hat{\beta}_0$ $\hat{\beta}_1$

Case Study IV: Wald procedures

(Similar test as in binary logistic regression)

- ▶ Hypotheses:

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

- ▶ Test statistic:

$$z = \frac{-0.2971}{0.0549} = -5.42 \sim N(0, 1) \text{ or } z^2 = 29.3 \sim \chi_1^2$$

$= (-5.42)^2$

- ▶ P-value < 0.0001

- ▶ Conclusion: Strong evidence that coefficient of log(Area) is not zero. Evidence that extinction probabilities are associated with island area.

- ▶ 95% CI for β_1 :

$$-0.2971 \pm 1.96(0.0549) = (-0.40, -0.19)$$

18/32

does not include 0

$$\hat{\beta}_1 \pm 1.96 (se(\hat{\beta}_1))$$

Case Study IV: Interpretation of β_1

$$\log a = b \rightarrow a = e^b$$

► Model:

$$\text{logit}(\pi) = \beta_0 + \beta_1 \log(x)$$

$$\Rightarrow \frac{\pi}{1 - \pi} = e^{\beta_0} e^{\beta_1 \log(x)} = e^{\beta_0} x^{\beta_1}$$

$$\log_{10} 10 = 1 \Rightarrow 10^1$$

$$\log_{10} 100 = 2 \Rightarrow 10^2 = 10(10)$$

► Interpretation: Hence, changing x by a factor of h , changes the odds by a multiplicative factor of h^{β_1} .

$$h = \frac{1}{2}$$

$$x \rightarrow xh$$

$$h = 2$$

$$\frac{e^{\beta_0} x^{\beta_1}}{e^{\beta_0}(1)}$$

19/32

Case Study IV: Interpretation of β_1

$$\frac{\pi}{1-\pi}$$

$\frac{1}{2}$

- ▶ **Example 1:** Halving island area changes odds by a factor of $0.5^{-0.2971} = 1.23$.

Therefore, the odds of extinction on a smaller island are 123% of the odds of extinction on an island double its size.

In other words, halving of area is associated with an increase in the odds of extinction by an estimated 23%.

An approximate 95% confidence interval for the percentage change in odds is 14% to 32%.

$\frac{2}{1}$

- ▶ **Example 2:** Doubling island area changes odds by a factor of $2^{-0.2971} = 0.81$.

Therefore, the odds of extinction for an at-risk species on a larger island are only 81% of the odds of extinction for such a species on an island half its size.

Case Study IV: Estimating probability of extinction

- Q: Estimate the probability of extinction for a species on the Ulkokrunni island.
- Fitted Model (M):

$$\text{logit}(\hat{\pi}_{M,i}) = -1.196 - 0.297 \log(\text{Area}_i)$$

- For Ulkokrunni island, $i = 1$ and $\text{Area}=185.5 \text{ km}^2$, then

$$\text{logit}(\hat{\pi}_{M,1}) = -1.196 - 0.297 \log(185.5) = *$$

Est. prob. $\hat{\pi}_{M,1} = \frac{e^*}{1+e^*}$

- Compared to the response proportion, $\bar{\pi}_{S,1} = \frac{5}{75} = 0.067$.

Obs. prob.

21/32

STA303/1004 - Class 11 R Markdown

February 8, 2018

22/32

Case Study IV: The Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case2101  
library(Sleuth3); krunnit = case2101  
str(krunnit)
```

```
## 'data.frame': 18 obs. of 4 variables:  
## $ Island : Factor w/ 18 levels "Hietakraasukka",...: 16 6 11 2 1 3 4 7 15 12  
## $ Area   : num  185.8 105.8 30.7 8.5 4.8 ...  
## $ AtRisk  : int  75 67 66 51 28 20 43 31 28 32 ...  
## $ Extinct: int  5 3 10 6 3 4 8 3 5 6 ...
```

x_i

m_i

y_i

Case Study IV: New variables

Get the data (from R library):

```
attach(krunnit); head(krunnit)
```

	Island	Area	AtRisk	Extinct
## 1	Ulkokrunni	185.8	75	5
## 2	Maakrunni	105.8	67	3
## 3	Ristikari	30.7	66	10
## 4	Isonkivenletto	8.5	51	6
## 5	Hietakraasukka	4.8	28	3
## 6	Kraasukka	4.5	20	4

m_i y_i

$$N_{\text{Extinct}} = m_i - y_i$$

$$75 - 5 = 70$$

$$67 - 3 = 64$$

:

$$20 - 4 = 16$$

$\log(\text{Area})$

```
logitpi<-log(Extinct/AtRisk/(1-(Extinct/AtRisk))) #observed logits
logarea<-log(Area) # log transformed Area
NExtinct<-AtRisk-Extinct
pis<-Extinct/AtRisk
```

$m_i - y_i$

we can compare: $\bar{\pi}_i$ with $\hat{\pi}_i$

: $\log \frac{\bar{\pi}_i}{1-\bar{\pi}_i}$ with $\log \frac{\hat{\pi}_i}{1-\hat{\pi}_i}$

$$\hat{\pi}_i$$

$$\bar{\pi}_i$$

$$\frac{s/75}{3/67}$$

$$\log \frac{\bar{\pi}_i}{1-\bar{\pi}_i}$$

$$(\frac{s/75}{3/67}) / (1 - (\frac{s/75}{3/67}))$$

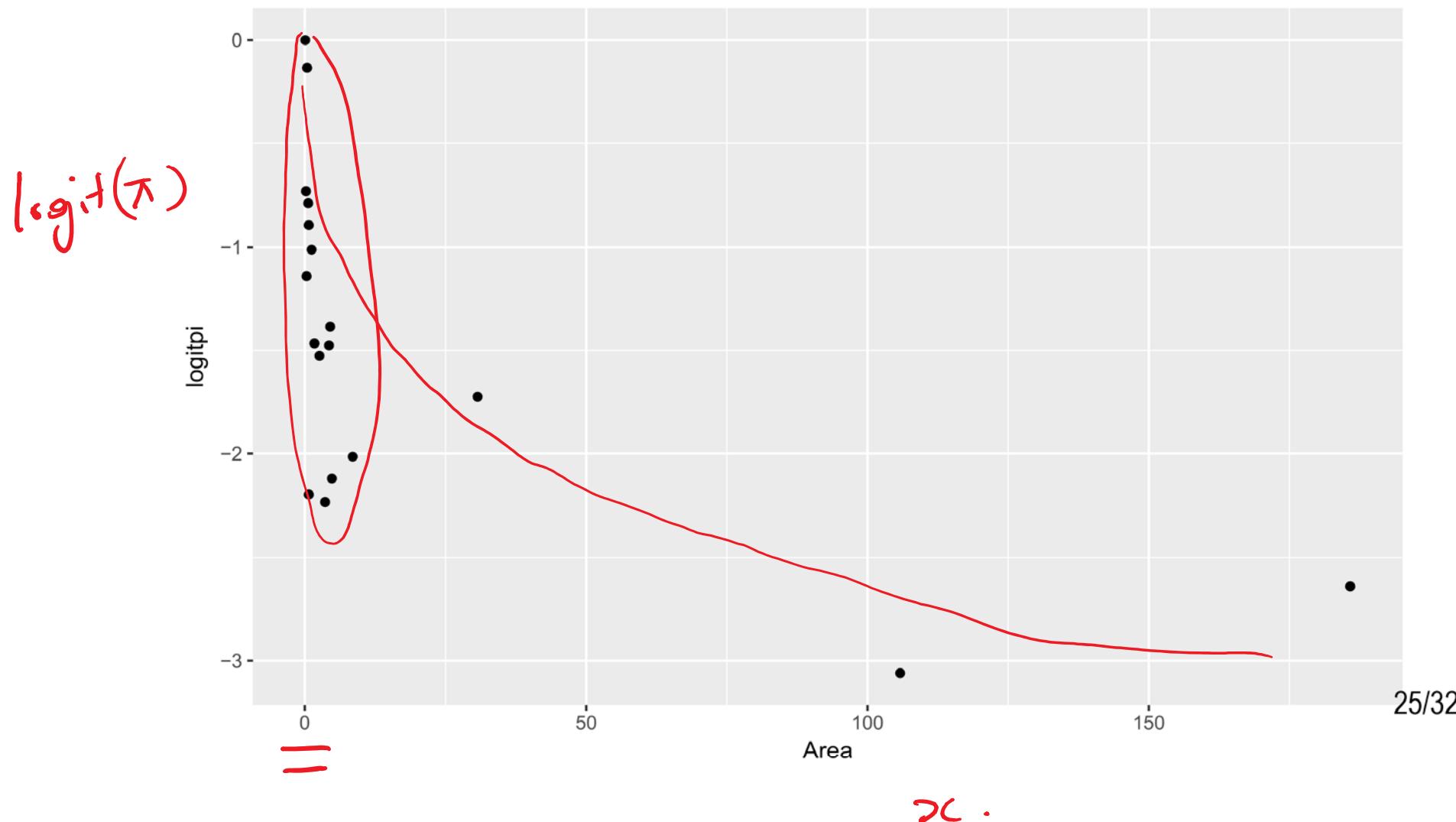
Empirical logits

$$- \log \left(\frac{\bar{\pi}_i}{1-\bar{\pi}_i} \right)$$

Eg, $\log \left(\frac{s/75}{1 - s/75} \right)$

Case Study IV: Visualizing the data

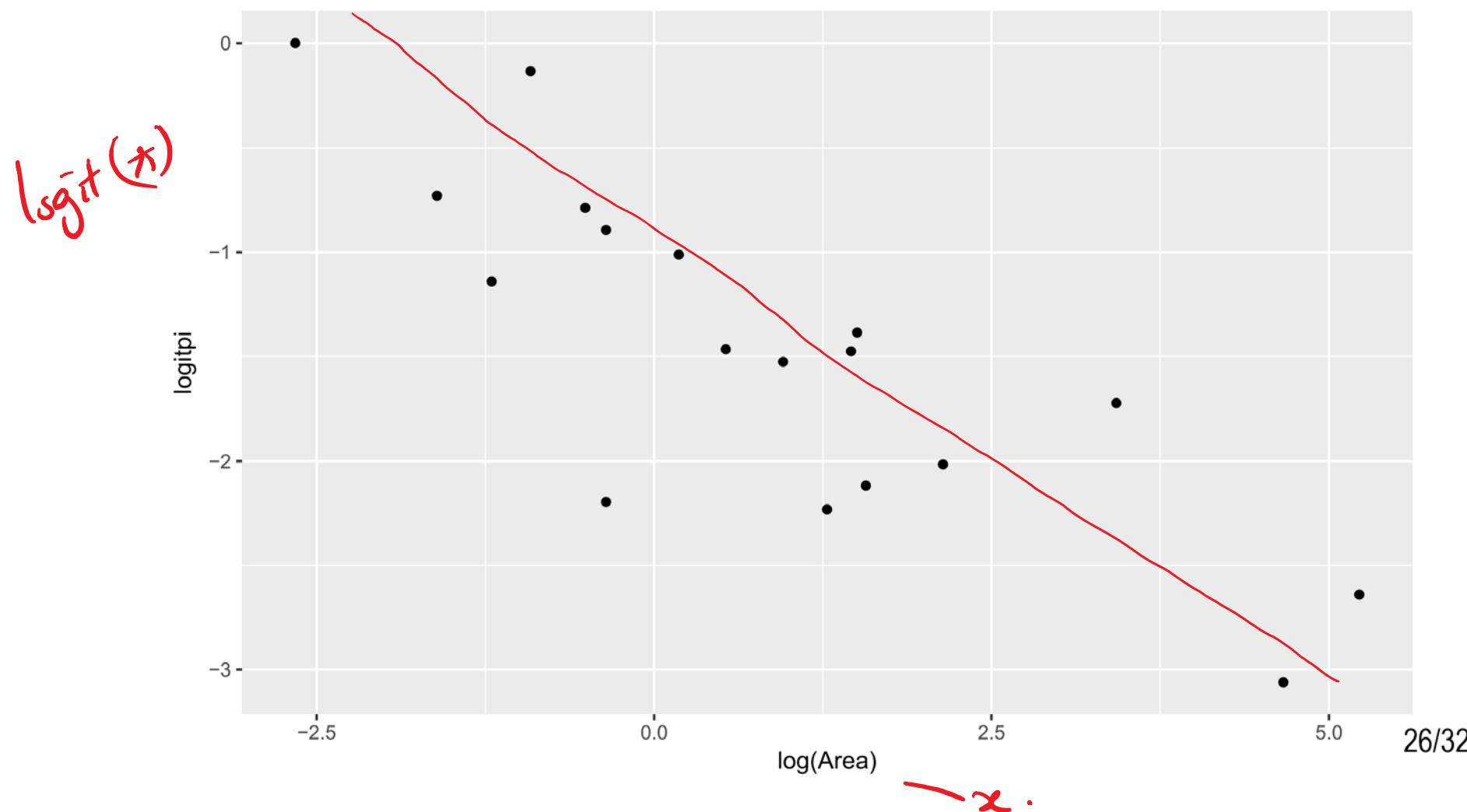
```
library(ggplot2)
ggplot(krunnit, aes(x=Area, y=logitpi)) + geom_point()
```



25/32

Case Study IV: Visualizing the data

```
ggplot(krunnit, aes(x=log(Area), y=logitpi)) + geom_point()
```



$$Y \sim X$$

Case Study IV: Logistic Model with logged explanatory variable

```
fitbl<-glm(cbind(Extinct, NExtinct)~log(Area), family=binomial, data=krunnit)
summary(fitbl)
```

of successes # of failures → m_i

Call:
 ## $glm(formula = cbind(Extinct, NExtinct) \sim log(Area), family = binomial,$
 ## data = krunnit)
 ##
 ## Deviance Residuals:
 ## Min 1Q Median 3Q Max
 ## -1.71726 -0.67722 0.09726 0.48365 1.49545
 ##
 ## Coefficients: $\hat{\beta}_j$ $se(\hat{\beta}_j)$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.19620	0.11845	-10.099	< 2e-16 ***
log(Area)	-0.29710	0.05485	-5.416	6.08e-08 ***

 ## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 ##
 ## (Dispersion parameter for binomial family taken to be 1)
 ##
 ## Null deviance: 45.338 on 17 degrees of freedom
 ## Residual deviance: 12.062 on 16 degrees of freedom
 ## AIC: 75.394
 ##
 ## Number of Fisher Scoring iterations: 4

$$Z^2 \sim \chi^2_1$$

$$(0.05485)^2 = 0.003$$

Case IV: Deviance test and Estimated Var-Cov of β

```
anova(fitbl, test="Chisq")  
  
## Analysis of Deviance Table  
##  
## Model: binomial, link: logit  
##  
## Response: cbind(Extinct, NExtinct)  
##  
## Terms added sequentially (first to last)  
##  
##  
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)  
## NULL             17    45.338  
## log(Area)     1   33.277      16    12.062 7.994e-09 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
print(vcov(fitbl))  
  
##          (Intercept)  log(Area)  
## (Intercept)  0.014029452 -0.002602237  
## log(Area)   -0.002602237 10.003008830
```

→ Used for Global LRT.

28/32

$$\text{var}(\hat{\beta}_1) = (\text{se}(\hat{\beta}_1))^2$$

Case IV: Wald tests in R

```
library(aod) # Analysis of Overdispersed Data  
wald.test(Sigma=vcov(fitbl), b=coef(fitbl), Terms=2)
```

```
## Wald test:  
## -----  
##
```

```
## Chi-squared test:
```

```
## X2 = 29.3, df = 1, P(> X2) = 6.1e-08
```

$$(-5.42)^2 = 29.3$$

$\text{var}(\hat{\beta}_j)$

$\hat{\beta}_j$

X^2

Case IV: Confidence Intervals for β 's

```
CL=cbind(bhat=coef(fitbl), confint.default(fitbl)) # 95% CI for betas  
CL
```

```
## bhat 2.5 % 97.5 %  
## (Intercept) -1.1961955 -1.4283454 -0.9640456  
## log(Area) -0.2971037 -0.4046132 -0.1895942
```

```
2^(CL) # doubling Area
```

```
## bhat 2.5 % 97.5 %  
## (Intercept) 0.4364247 0.3715568 0.5126174  
## log(Area) 0.8138847 0.7554388 0.8768524
```

```
.5^(CL) # halving Area
```

```
## bhat 2.5 % 97.5 %  
## (Intercept) 2.291346 2.691379 1.950773  
## log(Area) 1.228675 1.323734 1.140443
```

$$\hat{\beta}_1 \pm 1.96 \text{ se}(\hat{\beta}_1)$$

Case IV: Estimated probabilities of extinction per island

```

phats<-predict.glm(fitbl, type="response") # estimated probability of extinction
options(digits=4)
rbind(Extinct, NExtinct, pis, phats)

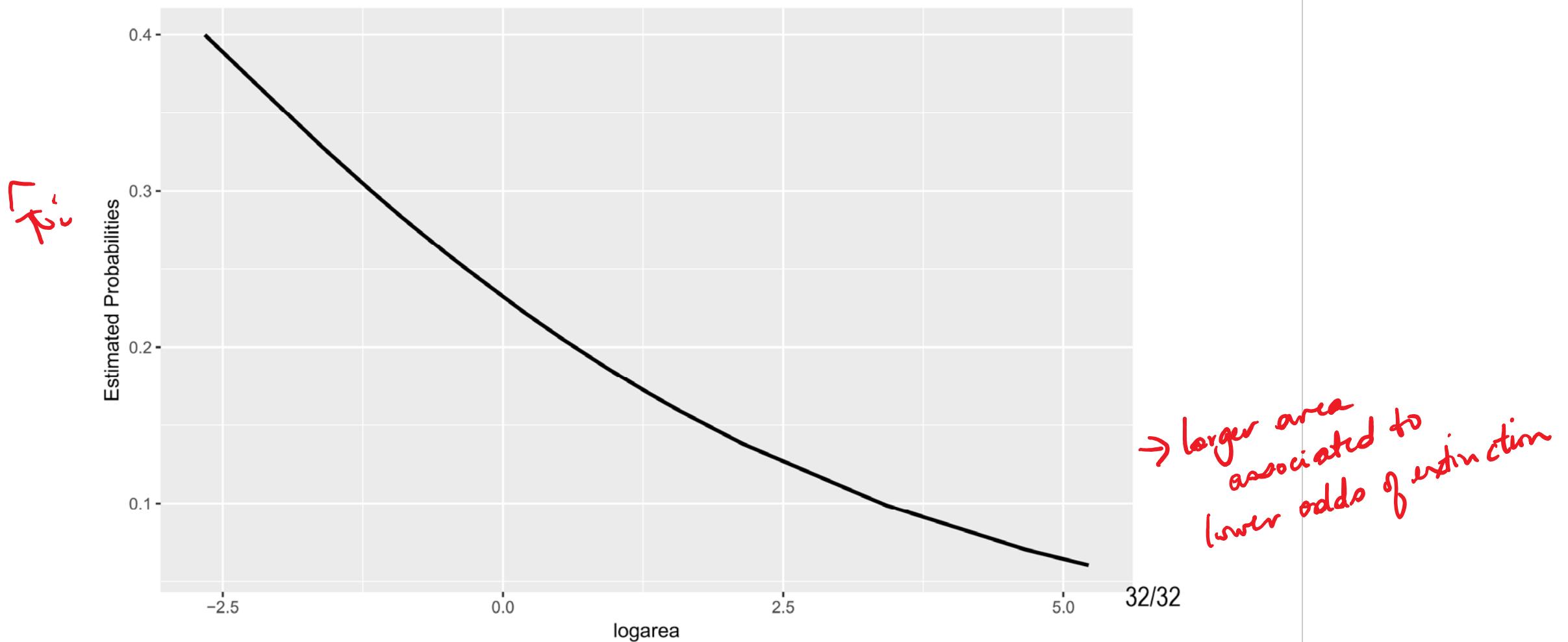
##          1      2      3      4      5      6      7
## Extinct 5.00000 3.00000 10.00000 6.00000 3.00000 4.000 8.0000
## NExtinct 70.00000 64.00000 56.00000 45.00000 25.00000 16.000 35.0000
## pis      0.06667 0.04478 0.15152 0.1176 0.1071 0.200 0.1860
## phats    0.06017 0.07036 0.09854 0.1380 0.1595 0.162 0.1639
##          8      9     10     11     12     13     14     15
## Extinct 3.00000 5.00000 6.00000 8.00000 2.00000 9.00000 5.00000 7.0000
## NExtinct 28.00000 23.00000 26.00000 22.00000 18.00000 22.00000 11.00000 8.0000
## pis      0.09677 0.1786 0.1875 0.2667 0.1000 0.2903 0.3125 0.4667
## phats    0.17125 0.1854 0.2052 0.2226 0.2516 0.2516 0.2603 0.2842
##          16     17     18
## Extinct 8.00000 13.00000 3.00000
## NExtinct 25.00000 27.00000 3.00000
## pis      0.2424 0.3250 0.5000
## phats    0.3019 0.3278 0.3998

```

$$\begin{aligned}
& \text{Observed} \\
& \underline{\text{pis} = \text{Extinct}} \\
& \text{Extinct} + \text{NExtinct} \\
& = \frac{\text{y}_i}{m_i} = \bar{\pi}_i \\
& \text{Estimated} \\
& \underline{\text{phats} \rightarrow \hat{\pi}_i}
\end{aligned}$$

Case IV Effect Plot

```
ggplot(krunnit,aes(x=logarea, y=phats))+ylab("Estimated Probabilities")+
  geom_line(size=1)
```



STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



February 13-15, 2018

1/39

Logistic Regression Diagnostics

STA 303/1002: Class 12- Logistic Regression

- ▶ What did we learn about Binary Logistic Regression?

- ▶ Underlying probability distribution of response: Bernoulli

▶ Outcome: Response variable, Y -binary

- ▶ Model:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = f(\mathbf{X}; \boldsymbol{\beta})$$

where $f(\mathbf{X}; \boldsymbol{\beta})$ is a linear function of the β 's

- ▶ Predictor variables, \mathbf{X} : categorical and/or continuous

- ▶ Estimation: MLE via Fisher scoring algorithm

- ▶ Interpretation of β 's: Hold other X 's constant, the odds of $Y=1$ change by factor of e^β .

- ▶ Estimate Odds, Odds ratio, $e^{\beta(a-b)}$

- ▶ Inference:

- ▶ Wald tests and confidence intervals

- ▶ Compare models: LRT: 1) $\underline{\beta}$, 2) $\underline{1}$, 3) $\underline{\text{Global}}$

y_i	i
0	1
1	2
0	:
0	0
1	:
0	N

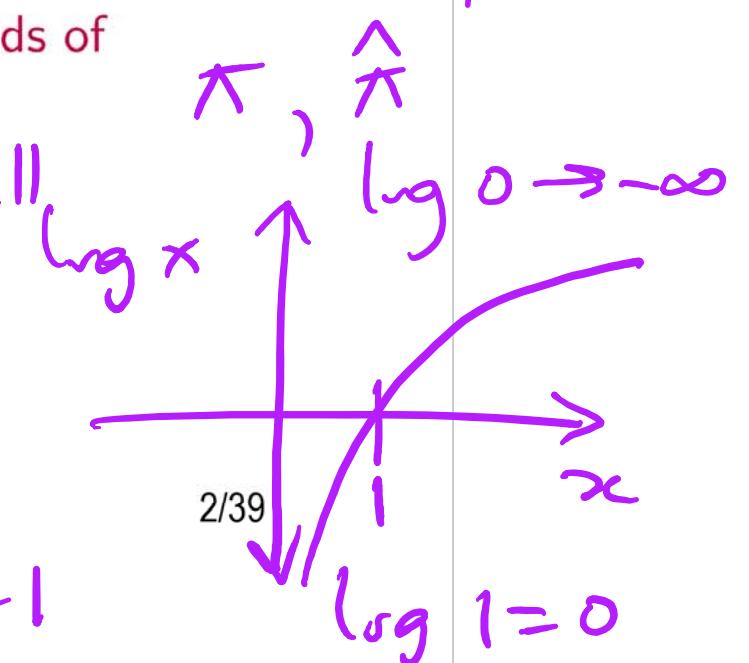
Logistic Regression Diagnostics

$$\pi = P(\text{"success"}) = \frac{y_i}{m_i = 1}$$

Observed | Estimated
 $y = \begin{cases} 0 \\ 1 \end{cases}, \hat{y}$
 $\pi = \begin{cases} 0 \\ 1 \end{cases}, 0 < \hat{\pi} < 1$

Do not compare

Fit vs Null



Binomial Logistic Regression

► What did we learn about Binomial Logistic Regression?

- Underlying probability distribution of response: Binomial
 - Outcome: Response variable, Y -count variable
- Model:

$$\log \left(\frac{\pi}{1 - \pi} \right) = f(\mathbf{X}; \boldsymbol{\beta})$$

$$\hat{\pi}_i = \frac{y_i}{m_i}$$

$$\hat{\pi}_i$$

where $f(\mathbf{X}; \boldsymbol{\beta})$ is a linear function of the β 's

- Estimate Odds, Odds Ratio and π
- Inference: Wald or LRT
- We can do more tests for model adequacy than in Binary logistic regression.

► Deviance GOF test: Fitted vs Saturated

- Quote of the week: "All models are wrong but some are useful." - *Unknown*.

Which is an example of a Generalized Linear Model?

$$E(Y|X)$$

✓ (a) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

✓ (b) $\mu[Y|X_1] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$

✓ (c) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

✓ (d) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 \log(X_2)$

✓ (e) $\mu[Y|X_1] = \beta_0 + \beta_1 10^{X_1}$

✗ (f) $\mu[Y|X_1, X_2] = \frac{\beta_0 + \beta_1 X_1}{\beta_0 + \beta_2 X_2}$

✗ (g) $\mu[Y|X_1] = \beta_0 + \exp(\beta_1 X_1)$

(h) $\mu[Y|X_1] = \beta_0 \exp(\beta_1 X_1)$ ← H·W

✓ (i) $\mu[Y|X_1, X_2] = \beta_1 X_1 \exp(\beta_0 + \beta_2 X_2)$

$E(Y|X) \sim \text{linear to } \beta's$

$E(Y|X) \sim \text{linear to } \beta's.$

→ General Linear Reg.

$g(\mu) = \mu$

$g(\mu) \neq \mu$ $g(\mu) = \underline{\underline{\log(\mu)}}$
 $g(\mu) = \underline{\underline{\ln(\mu)}}$

(g) $\log \mu = \log(\beta_0 + e^{\beta_1 X_1})$

Logistic Regression Diagnostics

(h) $\log \mu = (\log \beta_0) + \beta_1 X_1$

(i) $\log \mu = \log(\beta_1 X_1) +$

4/39

$= (\beta_0 + \log \beta_1 + \log X_1) + \beta_2 X_2$

Which is false?

- (i) A Logistic regression model is a Generalized Linear Model. — TRUE
- (ii) Logistic regression assumes that there is a linear relationship between logits and explanatory variables. $\logit(\pi) = \mathbf{X}\boldsymbol{\beta}$.
- (iii) Logistic regression describes population proportion or probability as a linear function of explanatory variables. π should be of f 's.
- (iv) Logistic regression is a nonlinear regression model. — TRUE.

Model Assumptions for Binomial Logistic Regression

1. Underlying probability model for response is Binomial.
 - ▶ Variance is not constant; is a function of the mean.
2. Observations are independent. (based on design of study)
- ③ The form of the model is correct
 - ▶ Linear relationship between logits and explanatory variables
 - ▶ All relevant variables are included; irrelevant ones excluded
4. Sample size is large enough for valid inference-tests and CIs.
(Recall large-sample properties of MLEs.)
 - ▶ Check for outliers.

$$E\left(\frac{y_i}{m_i}\right) = \pi_i$$

↓

$$\text{Var}\left(\frac{y_i}{m_i}\right) = \frac{\pi_i(1-\pi_i)}{m_i}$$

$$E(\varepsilon_i) = 0$$
$$\text{Var}(\varepsilon_i) = \sigma^2$$

What is the SATURATED Model?

- $\overbrace{y_i}^{\text{y}_i}$
- ▶ Observed response proportion:
$$\bar{\pi}_i = \frac{y_i}{m_i}$$

$$0 \leq y_i \leq m_i$$

$$0 \leq \bar{\pi}_i \leq 1$$
 - ▶ Observed or Empirical logits: (S-“saturated”)
$$\log\left(\frac{\bar{\pi}_{S,i}}{1 - \bar{\pi}_{S,i}}\right) = \log\left(\frac{y_i}{m_i - y_i}\right) = \log\left(\frac{y_i/m_i}{1 - y_i/m_i}\right)$$
 - ▶ Fits the model exactly with the data
 - ▶ Most general model possible for the data.

Which Models are often compared?

Consider one explanatory variable, X with n unique levels for the outcome, $Y \sim (Bin(m, \pi))$

- ▶ Saturated (FULL) Model: as many parameter coefficients as n

$$\text{logit}(\hat{\pi}) = \overbrace{\hat{\alpha}_0 + \hat{\alpha}_1 \mathbb{1}_1 + \cdots + \hat{\alpha}_{n-1} \mathbb{1}_{n-1}}^{n-1+1=n}$$

- ▶ Fitted (REDUCED) Model: nested within a FULL model; has $(p + 1)$ parameters

$$p+1, p < n$$

$$\text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 X$$

- ▶ NULL Model: Intercept only model

$$\text{logit}(\hat{\pi}) = \hat{\gamma}_0$$

Checking model adequacy: Form of the model

Deviance Goodness -Of -Fit (G-O-F) Test

- ▶ To check model adequacy in binomial logistic regression, we can use the Deviance Goodness -Of -Fit (G-O-F) Test.
- ▶ Analogous to GOF test for comparing 2 models in Linear Regression.

SATURATED .

- ▶ Form of hypotheses: H_0 : REDUCED model, H_a : FULL model
- ▶ The DEVIANCE GOF test compares the fitted model (M) to the saturated model (S).

$$H_0 : (\text{Fitted}) \text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$H_a : (\text{Saturated}) \text{logit}(\hat{\pi}) = \hat{\alpha}_0 + \hat{\alpha}_1 \mathbb{1}_1 + \cdots + \hat{\alpha}_{n-1} \mathbb{1}_{n-1}$$

Compared to Saturated model: Deviance G-O-F test

- ▶ Uses LRT
- ▶ Sometimes called “Drop-in-Deviance” test
- ▶ as extra-sum-of-squares tests; based on the deviance residual
- ▶ **Hypotheses:**

$$H_0: \text{logit}(\pi) = \alpha_0 + \alpha_1 X$$

(Fitted model fits data as well as Saturated model)

$$H_a: \text{logit}(\pi) = \beta_0 + \beta_1 \mathbb{1}_1 + \cdots + \beta_{n-1} \mathbb{1}_{n-1}$$

(Saturated model is better)

- ▶ **Test Statistic:**

$$\text{Deviance} = -2 \log \left(\frac{\mathcal{L}_R}{\mathcal{L}_F} \right) = -2 \log \left(\frac{\mathcal{L}_M}{\mathcal{L}_S} \right)$$

- ▶ Under H_0 , Deviance \sim Chi-square distribution with $n - (p + 1)$ df.
- ▶ **Warning:** This is an asymptotic approximation, so it works better if each $m_i > 5$.)

Calculating the Deviance test statistic

Recall underlying model of Y : $Y_i \sim \text{Binomial}(m_i, \pi_i)$

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}, \quad y_i = 0, 1, \dots, m_i$$

Hence the likelihood is:

$$\mathcal{L} = \prod_{i=1}^n \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}$$

$i = 0, 1, 2, \dots, n$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}$$

Calculating the Deviance test statistic

Then the log-likelihood is:

$$\log \mathcal{L} = \sum_{i=1}^n [y_i \log(\pi_i) + (m_i - y_i) \log(1 - \pi_i) + \underline{\log \binom{m_i}{y_i}}]$$

The deviance test statistic is based on a ratio of likelihoods.

$$\begin{aligned} \text{Deviance} &= -2 \log \frac{\mathcal{L}_M}{\mathcal{L}_S} \\ &= -2(\log \mathcal{L}_M - \log \mathcal{L}_S) \\ &= 2(\log \mathcal{L}_S - \log \mathcal{L}_M) \end{aligned}$$

- Q: A Saturated Model has $\text{Deviance} = 0$

$$\mathcal{L}_S = 1$$

Calculating the Deviance test statistic

$$\begin{aligned} \text{Deviance} &= 2(\log \mathcal{L}_S - \log \mathcal{L}_M) \\ &= 2 \sum_{i=1}^n \left[y_i \log \left(\frac{y_i}{m_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i} \right) + \log \left(\frac{m_i}{y_i} \right) \right. \\ &\quad \left. - y_i \log \left(\frac{\hat{y}_i}{m_i} \right) - (m_i - y_i) \log \left(\frac{m_i - \hat{y}_i}{m_i} \right) - \log \left(\frac{m_i}{\hat{y}_i} \right) \right] \\ &= 2 \sum_{i=1}^n \left[y_i \log(y_i) + (m_i - y_i) \log(m_i - y_i) \right. \\ &\quad \left. - y_i \log(\hat{y}_i) - (m_i - y_i) \log(m_i - \hat{y}_i) \right] \\ &= 2 \sum_{i=1}^n \left[y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + \cancel{(m_i - y_i)} \log \left(\frac{\cancel{m_i - y_i}}{\underline{m_i - \hat{y}_i}} \right) \right] \end{aligned}$$

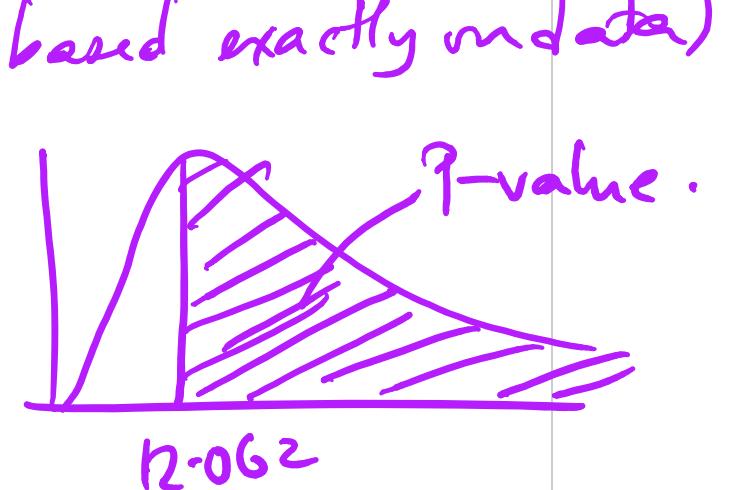
Case Study IV Exercise: Using Deviance

Using R output,

Q: Determine whether a saturated model is an improvement over the simpler model with linear function of $\log(\text{Area})$.

(In R, we get deviance of a model by using deviance('fittedmodel'))

- ▶ Hypotheses: $H_0: \text{Fitted} \Rightarrow \text{logit}(z) = X\beta$.
 $H_a: \text{Saturated} \quad \text{logit}(z) = (\text{based exactly on data})$.
- ▶ Test Statistic: Deviance=12.062
- ▶ In R: Residual deviance
- ▶ Distribution of TS: $\sim \chi^2_{16}$
- ▶ P-value: $P(\chi^2_{16} \geq 12.062) = 0.74$
- ▶ Conclusion: The data are consistent with H_0 ; the simpler model with linear function of $\log(\text{Area})$ is adequate (fits as well as the saturated model).



Binomial Logistic Regression: Interpreting Deviance

- ▶ Smaller deviance leads to larger p -value and vice versa.
- ▶ Large p -values means:
 - ▶ Fitted model is adequate, OR
 - ▶ Test is not powerful enough to detect inadequacies
- ▶ Small p -values means:
 - ▶ Fitted model is not adequate; consider a more complex model with more explanatory variables or higher order terms and so on, OR
 - ▶ Response distribution is not adequately modelled by the Binomial distribution, OR
 - ▶ There are severe outliers.

Eg, Poisson .

Can we do a Deviance GOF test in Binary case?

In Binary logistic regression case, $m_i = 1$ for all i , and $y_i = \begin{cases} 0 \\ 1 \end{cases}$

Then deviance becomes:

$$\begin{aligned} \text{Deviance} &= 2 \sum_{i=1}^n [y_i \log(y_i) + (1 - y_i) \log(1 - y_i) \\ &\quad - y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)] \end{aligned}$$

sat, \mathcal{L}_S \nearrow Fitted, \mathcal{L}_M .

$$= 2 \sum_{i=1}^n [-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)].$$

Notice that the terms that came from the saturated model, $\log \mathcal{L}_S$ are gone, so deviance is no longer useful to compare \mathcal{L}_M with \mathcal{L}_S .

Model assessment in Binomial Logistic Regression

- ▶ Is linear relationship appropriate?
 - ▶ Plot observed logit versus quantitative explanatory variable
- ▶ Is the form of the model correct?
 - ▶ Use Wald or LRT tests
- ▶ Is saturated model better than fitted model?
 - ▶ Deviance GOF test
- ▶ Are there outliers?
 - ▶ Examine standardized residuals: Pearson and Deviance Residuals
- ▶ Consider other model fit statistics: AIC, BIC
- ▶ Other issues/concerns in model fitting

Residuals: Pearson and Deviance

- ▶ Response (raw) residuals: (*observed – fitted*) proportion

$$\hat{\pi}_{S,i} - \hat{\pi}_{M,i} = \frac{y_i}{m_i} - \hat{\pi}_{M,i}$$

- ▶ Standardized residuals:

- (1) Pearson Residuals: uses estimate of s.d. of Y (in denominator)

$$P_{res,i} = \frac{y_i - m_i \hat{\pi}_{M,i}}{\sqrt{m_i \hat{\pi}_{M,i} (1 - \hat{\pi}_{M,i})}}$$

- (2) Deviance Residuals: defined so that the sum of the squares of the residuals is the deviance

$$D_{res,i} = \text{sign}(y_i - m_i \hat{\pi}_{M,i}) \times \sqrt{2 \left\{ y_i \log \left(\frac{y_i}{m_i \hat{\pi}_{M,i}} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i - m_i \hat{\pi}_{M,i}} \right) \right\}}$$

Response, Pearson and Deviance Residuals in R

- ▶ Response residuals

```
residuals(fitbl, type="response")
```

- ▶ Pearson residuals

```
residuals(fitbl, type="pearson")
```

- ▶ Deviance residuals

```
residuals(fitbl, type="deviance")
```

Case Study IV Example: Were there outliers in the data?

	Pearson, $P_{res,i}$	Deviance, $D_{res,i}$
Asymptotic Dist. R code	$N(0, 1)$ pearson	$N(0, 1)$ deviance
Possible outlier if	$ P_{res,i} > 2$	$ D_{res,i} > 2$
Outlier if	$ P_{res,i} > 3$	$ D_{res,i} > 3$
Under small n	D_{res} closer to $N(0, 1)$ than P_{res}	
$\hat{\pi}$ close to 0 or 1	P_{res} are unstable; related to instability of Wald	

- ▶ Results: Both are $< |2|$, so no outliers

Other Model Fit Statistics

- ▶ Useful for comparing models with same response and same data
- ▶ Two popular fit statistics: AIC and BIC; combines log-likelihood with a penalty
 1. Akaike's Information Criterion (AIC)

$$AIC = -2 \log \mathcal{L} + 2(p + 1)$$

- 2. Schwarz's (Bayesian Information) Criterion (BIC)

$$BIC = -2 \log \mathcal{L} + (p + 1) \log N$$

where

- ▶ p -number of explanatory variables, and
 - ▶ $N = \sum_{i=1}^n m_i$.
- ▶ Example: see AIC, BIC for Case IV model

Problems and Complications common to Linear and Logistic Regression

- ▶ *Extrapolation*- don't make inferences/predictions outside range of observed data; model may no longer be appropriate.
- ▶ *Multicollinearity*- highly correlated explanatory variables; difficult to assess individual effects on response. Consequences include:
 - ▶ Unstable fitted equation
 - ▶ Coefficient that should be statistically significant is not
 - ▶ Coefficient may have the wrong sign
 - ▶ Sometimes, large s.e. of $\hat{\beta}$
 - ▶ Sometimes numerical procedure to find MLEs does not converge

Problems and Complications common to Linear and Logistic Regression

- ▶ *Extrapolation*- don't make inferences/predictions outside range of observed data; model may no longer be appropriate.
- ▶ *Multicollinearity*- highly correlated explanatory variables; difficult to assess individual effects on response. Consequences include:
 - ▶ Unstable fitted equation
 - ▶ Coefficient that should be statistically significant is not
 - ▶ Coefficient may have the wrong sign
 - ▶ Sometimes, large s.e. of $\hat{\beta}$
 - ▶ Sometimes numerical procedure to find MLEs does not converge

Problems and Complications common to Linear and Logistic Regression

- ▶ *Influential points*- an observation is influential if its removal substantially changes estimated coefficients (such as, fitted $\hat{\beta}$'s, deviance)
- ▶ *Model Building*- choosing explanatory variables and their forms (eg. polynomial terms, interaction and transformations) tend to overfit the data; should build model on training data and test on test data (cross validation).

Problems and Complications common to Linear and Logistic Regression

- ▶ *Influential points*- an observation is influential if its removal substantially changes estimated coefficients (such as, fitted $\hat{\beta}$'s, deviance)
- ▶ *Model Building*- choosing explanatory variables and their forms (eg. polynomial terms, interaction and transformations) tend to overfit the data; should build model on training data and test on test data (cross validation).

Two problems specific to Logistic Regression

1. Extra-binomial variation

- ▶ variance of Y_i greater than $m_i\pi_i(1 - \pi_i)$
- ▶ also called “over dispersion”
- ▶ does not bias $\hat{\beta}$'s but s.e. of $\hat{\beta}$'s will be too small
(too small p -values, too narrow CIs)

SOLUTION: add one more parameter to the model, ψ - dispersion parameter. Then $\text{Var}(Y_i) = \psi m_i\pi_i(1 - \pi_i)$.

Two problems specific to logistic regression

2. Complete and Quasi-complete separation

- ▶ *Complete separation:*

- ▶ one or a linear combination of explanatory variables perfectly predict whether $Y = 1$ or $Y = 0$
- ▶ In Binary response, when $y_i = 1$, $\hat{y}_i = 1$, then $\sum_{i=1}^n \{y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\} = 0$.
- ▶ MLE's cannot be computed

- ▶ *Quasi-complete separation:*

- ▶ explanatory variables predict $Y = 1$ or $Y = 0$ almost perfectly (just a few points wrong)
- ▶ MLE's are numerically unstable

SOLUTION: simplify the model. Other options- penalized maximum likelihood, exact logistic regression, bayesian methods

Using Logistic Regression for Classification

- ▶ **Want:** predict outcome as

$$y^* | (x_1^*, x_2^*, \dots, x_p^*) = \begin{cases} 1 \\ 0 \end{cases}$$

- ▶ **Do:** calculate $\hat{\pi}_M^*$ - the estimated probability that $y^* = 1$ based on the fitted model given $X_1 = x_1^*, X_2 = x_2^*, \dots, X_p = x_p^*$.
From this we want to predict that

$$y^* = \begin{cases} 1 & \text{if } \hat{\pi}_M^* \text{ is large} \\ 0 & \text{if } \hat{\pi}_M^* \text{ is small} \end{cases}$$

- ▶ **Need:** choose a cut-off probability to distinguish between large and small.

Classification: Approaches to choosing a threshold

Approach 1 - Set cut-off probability as 0.5

- ▶ If $\hat{\pi}_M^* > 0.5$, classify y^* as 1
- ▶ Useful if there are equal numbers of 1's and 0's
- ▶ Useful if false negatives and false positives are equally bad.

Classification: Approaches to choosing a threshold

Approach 2- Find “best” cut-off probability from data.

- ▶ Try different cut-offs and see which gives fewest incorrect classifications
- ▶ Useful if proportions of 1's and 0's in data reflect their relative proportions in the population
- ▶ Likely to overestimate the proportions of correct predictions that model makes. Then, one should assess model correct classification rates on different data than was used to fit the model.

STA303/1004 - Class 12 R Markdown

February 13, 2018

Case Study IV: The Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case2101
library(Sleuth3); krunnit = case2101
str(krunnit)

## 'data.frame':    18 obs. of  4 variables:
## $ Island : Factor w/ 18 levels "Hietakraasukka",...: 16 6 11 2 1 3 4 7 15 12
## $ Area   : num  185.8 105.8 30.7 8.5 4.8 ...
## $ AtRisk : int  75 67 66 51 28 20 43 31 28 32 ...
## $ Extinct: int  5 3 10 6 3 4 8 3 5 6 ...
```

Case Study IV: Logistic Model with logged explanatory variable

```
fitbl<-glm(cbind(Extinct,NExtinct)~log(Area), family=binomial, data=krunnit)
summary(fitbl)
```

```
##
## Call:
## glm(formula = cbind(Extinct, NExtinct) ~ log(Area), family = binomial,
##      data = krunnit)
##
## Deviance Residuals:
##       Min        1Q     Median        3Q       Max
## -1.71726  -0.67722   0.09726   0.48365   1.49545
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.19620   0.11845 -10.099 < 2e-16 ***
## log(Area)   -0.29710   0.05485  -5.416 6.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Dispersion parameter for binomial family taken to be 1
##
## Null deviance: 45.338 on 17 degrees of freedom
## Residual deviance: 12.062 on 16 degrees of freedom
## AIC: 75.394
##
## Number of Fisher Scoring iterations: 4
```

In R:
deviance(fitbl).

= 12.062

Case IV: Deviance test and Estimated Var-Cov of β

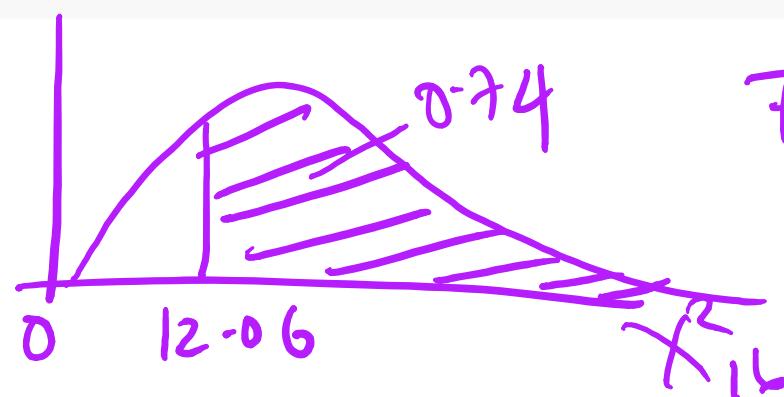
```
anova(fitbl, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Extinct, NExtinct)
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL              17     45.338
## log(Area)    1   33.277      16    12.062 7.994e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
1-pchisq(12.062, 16)
```

```
## [1] 0.7397009
```

Global LRT:
(NULL vs Fitted)



Fitted vs Std.

Case IV: Estimated probabilities of extinction per island

```
phats<-predict.glm(fitbl, type="response") # estimated probability of extinction
options(digits=4)
rbind(Extinct, NExtinct, pis,phats)

##          1      2      3      4      5      6      7
## Extinct 5.00000 3.00000 10.00000 6.00000 3.00000 4.000 8.0000
## NExtinct 70.00000 64.00000 56.00000 45.00000 25.00000 16.000 35.0000
## pis      0.06667 0.04478 0.15152 0.1176 0.1071 0.200 0.1860
## phats    0.06017 0.07036 0.09854 0.1380 0.1595 0.162 0.1639
##          8      9     10     11     12     13     14     15
## Extinct 3.00000 5.0000 6.0000 8.0000 2.0000 9.0000 5.0000 7.0000
## NExtinct 28.00000 23.0000 26.0000 22.0000 18.0000 22.0000 11.0000 8.0000
## pis      0.09677 0.1786 0.1875 0.2667 0.1000 0.2903 0.3125 0.4667
## phats    0.17125 0.1854 0.2052 0.2226 0.2516 0.2516 0.2603 0.2842
##          16     17     18
## Extinct 8.0000 13.0000 3.0000
## NExtinct 25.0000 27.0000 3.0000
## pis      0.2424 0.3250 0.5000
## phats    0.3019 0.3278 0.3998
```

Case IV Fit Statistics

```
AIC(fitbl)
```

```
## [1] 75.39
```

```
BIC(fitbl)
```

```
## [1] 77.17
```

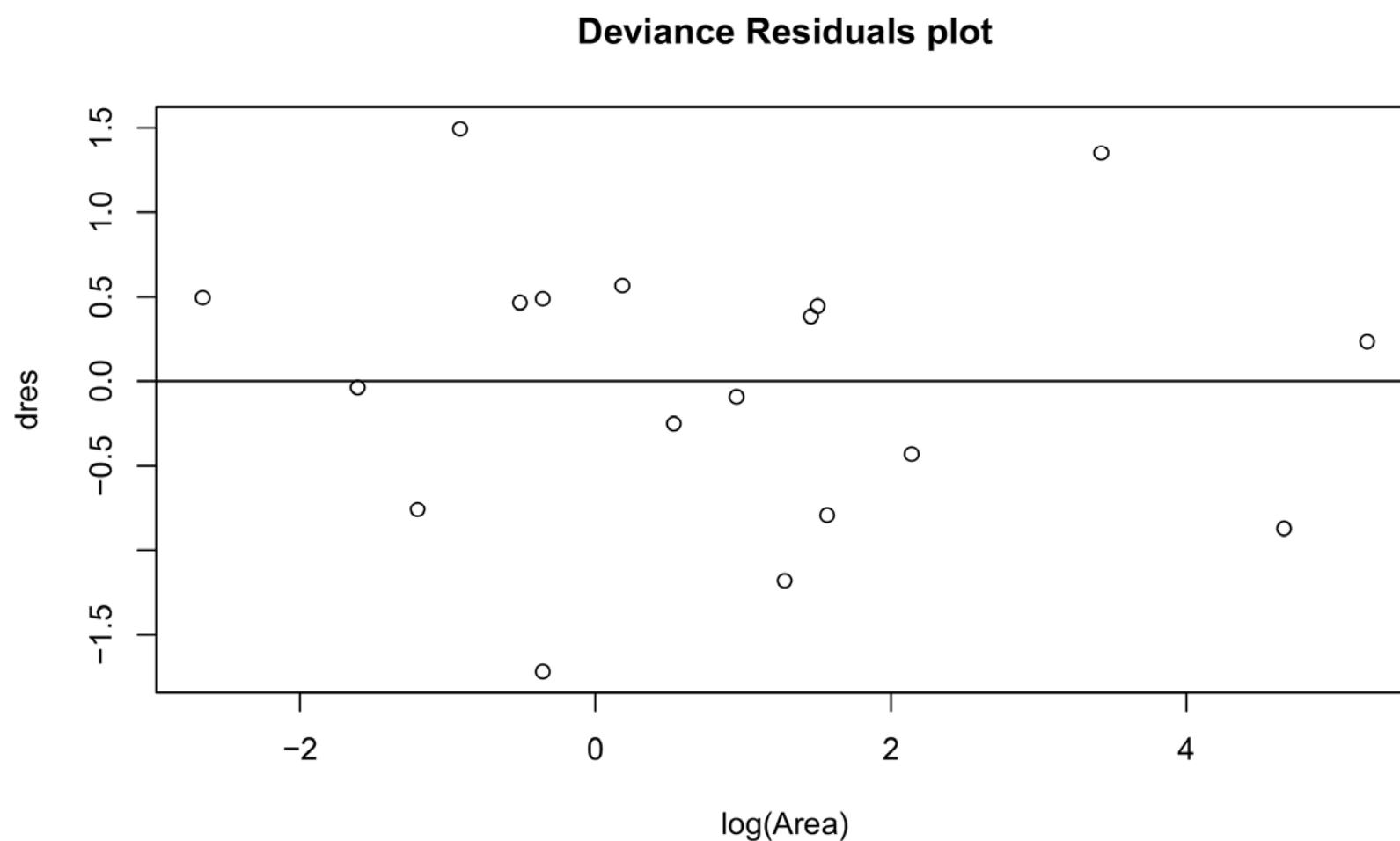
Case IV Residuals

```
rres<-residuals(fitbl, type=c("response"))
pres<-residuals(fitbl, type=c("pearson"))
dres<-residuals(fitbl, type=c("deviance"))
rbind(pis,phats,rres, pres,dres)
```

π_i — ## pis 0.066667 0.04478 0.15152 0.11765 0.10714 0.20000 0.18605 0.09677
 $\hat{\pi}_i$ — ## phats 0.060173 0.07036 0.09854 0.13800 0.15946 0.16205 0.16389 0.17125
rres 0.006493 -0.02558 0.05298 -0.02035 -0.05232 0.03795 0.02216 -0.07448
pres 0.236464 -0.81883 1.44400 -0.42139 -0.75619 0.46058 0.39247 -1.10075
dres 0.232656 -0.87369 1.34958 -0.43071 -0.79584 0.44746 0.38577 -1.18097
9 10 11 12 13 14 15 16
pis 0.178571 0.18750 0.26667 0.1000 0.29032 0.3125 0.4667 0.24242
phats 0.185415 0.20524 0.22264 0.2516 0.25158 0.2603 0.2842 0.30185
rres -0.006844 -0.01774 0.04403 -0.1516 0.03875 0.0522 0.1825 -0.05943
pres -0.093181 -0.24850 0.57969 -1.5622 0.49717 0.4759 1.5673 -0.74367
dres -0.093632 -0.25127 0.56727 -1.7173 0.48934 0.4666 1.4954 -0.75939
17 18
pis 0.325000 0.5000
phats 0.327828 0.3998
rres -0.002828 0.1002
pres -0.038101 0.5008
dres -0.038129 0.4957

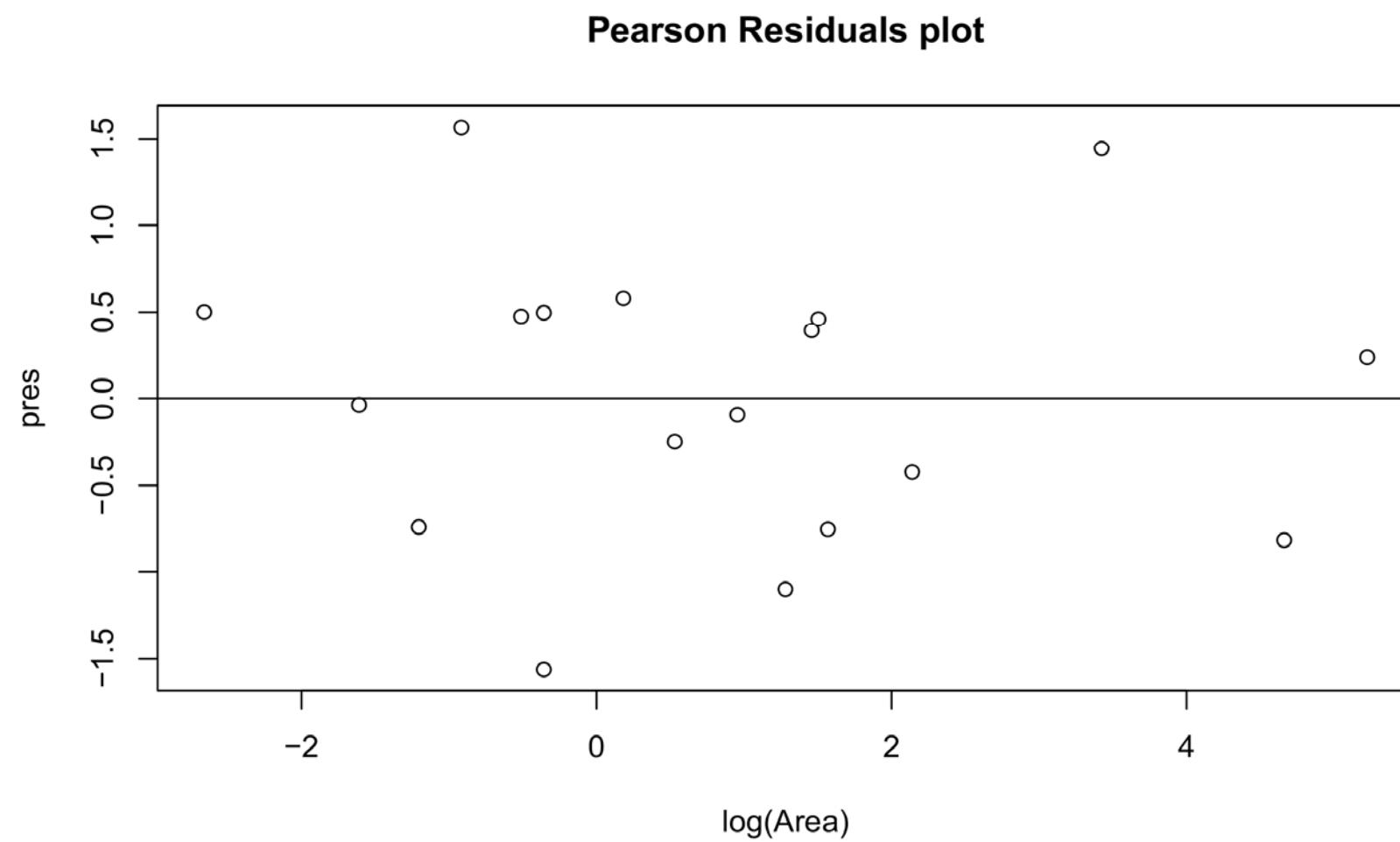
Case IV Residuals Plot

```
plot(log(Area), dres, main="Deviance Residuals plot")
abline(h=0)
```



Case IV Residuals Plot

```
plot(log(Area), pres, main="Pearson Residuals plot")
abline(h=0)
```



STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



February 13-15, 2018

1/41

STA 303/1002: Class 12- Logistic Regression

- ▶ What did we learn about Binary Logistic Regression?

- ▶ Underlying probability distribution of response: Bernoulli
 - ▶ Outcome: Response variable, Y -binary
- ▶ Model:

$$\log \left(\frac{\pi}{1 - \pi} \right) = f(\mathbf{X}; \boldsymbol{\beta})$$

where $f(\mathbf{X}; \boldsymbol{\beta})$ is a linear function of the β 's

- ▶ Predictor variables, \mathbf{X} : categorical and/or continuous
- ▶ Estimation: MLE via Fisher scoring algorithm
- ▶ Interpretation of β 's: Hold other X 's constant, the odds of $Y=1$ change by factor of e^{β} .
- ▶ Estimate Odds, Odds ratio, $e^{\beta(a-b)}$
- ▶ Inference:
 - ▶ Wald tests and confidence intervals
 - ▶ Compare models: LRT: 1) $> 1 \beta$, 2) 1β , 3) Global

Binomial Logistic Regression

- ▶ What did we learn about Binomial Logistic Regression?
 - ▶ Underlying probability distribution of response: Binomial
 - ▶ Outcome: Response variable, Y -count variable
 - ▶ Model:
$$\log \left(\frac{\pi}{1 - \pi} \right) = f(\mathbf{X}; \boldsymbol{\beta})$$
where $f(\mathbf{X}; \boldsymbol{\beta})$ is a linear function of the β 's
 - ▶ Estimate Odds, Odds Ratio and π
 - ▶ Inference: Wald or LRT
 - ▶ We can do more tests for model adequacy than in Binary logistic regression.
- ▶ Deviance GOF test: Fitted vs Saturated
- ▶ Quote of the week: “All models are wrong but some are useful.” - *Unknown*.

Which is an example of a Generalized Linear Model?

- (a) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- (b) $\mu[Y|X_1] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$
- (c) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- (d) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 \log(X_2)$
- (e) $\mu[Y|X_1] = \beta_0 + \beta_1 10^{X_1}$
- (f) $\mu[Y|X_1, X_2] = \frac{\beta_0 + \beta_1 X_1}{\beta_0 + \beta_2 X_2}$
- (g) $\mu[Y|X_1] = \beta_0 + \exp(\beta_1 X_1)$
- (h) $\mu[Y|X_1] = \beta_0 \exp(\beta_1 X_1)$
- (i) $\mu[Y|X_1, X_2] = \beta_1 X_1 \exp(\beta_0 + \beta_2 X_2)$

Which is false?

- (i) A Logistic regression model is a Generalized Linear Model.
- (ii) Logistic regression assumes that there is a linear relationship between logits and explanatory variables.
- (iii) Logistic regression describes population proportion or probability as a linear function of explanatory variables.
- (iv) Logistic regression is a nonlinear regression model.

Model Assumptions for Binomial Logistic Regression

1. Underlying probability model for response is Binomial.
 - ▶ Variance is not constant; is a function of the mean.
2. Observations are independent.
3. The form of the model is correct
 - ▶ Linear relationship between logits and explanatory variables
 - ▶ All relevant variables are included; irrelevant ones excluded
4. Sample size is large enough for valid inference-tests and CIs.
(Recall large-sample properties of MLEs.)
 - ▶ Check for outliers.

What is the SATURATED Model?

- ▶ Observed response proportion:

$$\bar{\pi}_i = \frac{y_i}{m_i}$$

- ▶ Observed or Empirical logits: (S-“saturated”)

$$\log\left(\frac{\bar{\pi}_{S,i}}{1 - \bar{\pi}_{S,i}}\right) = \log\left(\frac{y_i}{m_i - y_i}\right)$$

- ▶ Fits the model exactly with the data
- ▶ Most general model possible for the data.

Which Models are often compared?

Consider one explanatory variable, X with n unique levels for the outcome, $Y \sim (Bin(m, \pi))$

- ▶ Saturated (FULL) Model: as many parameter coefficients as n

$$\text{logit}(\hat{\pi}) = \hat{\alpha}_0 + \hat{\alpha}_1 \mathbb{1}_1 + \cdots + \hat{\alpha}_{n-1} \mathbb{1}_{n-1}$$

- ▶ Fitted (REDUCED) Model: nested within a FULL model; has $(p + 1)$ parameters

$$\text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 X$$

- ▶ NULL Model: Intercept only model

$$\text{logit}(\hat{\pi}) = \hat{\gamma}_0$$

Checking model adequacy: Form of the model

Deviance Goodness -Of -Fit (G-O-F) Test

- ▶ To check model adequacy in binomial logistic regression, we can use the Deviance Goodness -Of -Fit (G-O-F) Test.
- ▶ Analogous to GOF test for comparing 2 models in Linear Regression.
- ▶ Form of hypotheses: H_0 : REDUCED model, H_a : FULL model
- ▶ The DEVIANCE GOF test compares the fitted model (M) to the saturated model (S).

$$H_0 : (\text{Fitted}) \text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$H_a : (\text{Saturated}) \text{logit}(\hat{\pi}) = \hat{\alpha}_0 + \hat{\alpha}_1 \mathbb{1}_1 + \cdots + \hat{\alpha}_{n-1} \mathbb{1}_{n-1}$$

Compared to Saturated model: Deviance G-O-F test

- ▶ Uses LRT
- ▶ Sometimes called “Drop-in-Deviance” test
- ▶ as extra-sum-of-squares tests; based on the deviance residual
- ▶ Hypotheses:

$$H_0: \text{logit}(\pi) = \alpha_0 + \alpha_1 X \quad (\text{Fitted model fits data as well as Saturated model})$$

$$H_a: \text{logit}(\pi) = \beta_0 + \beta_1 \mathbb{1}_1 + \cdots + \beta_{n-1} \mathbb{1}_{n-1} \quad (\text{Saturated model is better})$$

- ▶ Test Statistic:

$$\text{Deviance} = -2 \log \left(\frac{\mathcal{L}_R}{\mathcal{L}_F} \right) = -2 \log \left(\frac{\mathcal{L}_M}{\mathcal{L}_S} \right)$$

- ▶ Under H_0 , Deviance \sim Chi-square distribution with $\underline{n - (p + 1)}$ df.
- ▶ Warning: This is an asymptotic approximation, so it works better if each $m_i > 5$.)

Calculating the Deviance test statistic

Recall underlying model of Y : $Y_i \sim \text{Binomial}(m_i, \pi_i)$

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}, \quad y_i = 0, 1, \dots, m_i$$

Hence the likelihood is:

$$\mathcal{L} = \prod_{i=1}^n \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}$$

Calculating the Deviance test statistic

Then the log-likelihood is:

$$\log \mathcal{L} = \sum_{i=1}^n [y_i \log(\hat{\pi}_i) + (m_i - y_i) \log(1 - \hat{\pi}_i) + \log \binom{m_i}{y_i}]$$

The deviance test statistic is based on a ratio of likelihoods.

$$\begin{aligned} \text{Deviance} &= -2 \log \frac{\mathcal{L}_M}{\mathcal{L}_S} \\ &= -2(\log \mathcal{L}_M - \log \mathcal{L}_S) \\ &= 2(\log \mathcal{L}_S - \log \mathcal{L}_M) \end{aligned}$$

- Q: A Saturated Model has *Deviance* =

Calculating the Deviance test statistic

$$\begin{aligned} \text{Deviance} &= 2(\log \mathcal{L}_S - \log \mathcal{L}_M) \\ &= 2 \sum_{i=1}^n \left[y_i \log \left(\frac{y_i}{m_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i} \right) + \log \left(\frac{m_i}{y_i} \right) \right. \\ &\quad \left. - y_i \log \left(\frac{\hat{y}_i}{m_i} \right) - (m_i - y_i) \log \left(\frac{m_i - \hat{y}_i}{m_i} \right) - \log \left(\frac{m_i}{\hat{y}_i} \right) \right] \\ &= 2 \sum_{i=1}^n \left[y_i \log(y_i) + (m_i - y_i) \log(m_i - y_i) \right. \\ &\quad \left. - y_i \log(\hat{y}_i) - (m_i - y_i) \log(m_i - \hat{y}_i) \right] \\ &= 2 \sum_{i=1}^n \left[y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i - \hat{y}_i} \right) \right] \end{aligned}$$

Case Study IV Exercise: Using Deviance

Using R output,

Q: Determine whether a saturated model is an improvement over the simpler model with linear function of $\log(\text{Area})$.

(In R, we get deviance of a model by using `deviance('fittedmodel')`)

- ▶ Hypotheses:
- ▶ Test Statistic: Deviance=12.062
- ▶ Distribution of TS:
- ▶ P-value:
- ▶ Conclusion: The data are consistent with H_0 ; the simpler model with linear function of $\log(\text{Area})$ is adequate (fits as well as the saturated model).

Binomial Logistic Regression: Interpreting Deviance

- ▶ Smaller deviance leads to larger p -value and vice versa.
- ▶ Large p -values means:
 - ▶ Fitted model is adequate, OR
 - ▶ Test is not powerful enough to detect inadequacies
- ▶ Small p -values means:
 - ▶ Fitted model is not adequate; consider a more complex model with more explanatory variables or higher order terms and so on, OR
 - ▶ Response distribution is not adequately modelled by the Binomial distribution, OR
 - ▶ There are severe outliers.

Can we do a Deviance GOF test in Binary case?

In Binary logistic regression case, $m_i = 1$ for all i , and $y_i = \begin{cases} 0 \\ 1 \end{cases}$

Then deviance becomes:

$$\begin{aligned} \text{Deviance} &= 2 \sum_{i=1}^n [y_i \log(y_i) + (1 - y_i) \log(1 - y_i) \\ &\quad - y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)] \\ &= 2 \sum_{i=1}^n [-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)]. \end{aligned}$$

Notice that the terms that came from the saturated model, $\log \mathcal{L}_S$ are gone, so deviance is no longer useful to compare \mathcal{L}_M with \mathcal{L}_S .

Model assessment in Binomial Logistic Regression

- ▶ Is linear relationship appropriate?
 - ▶ Plot observed logit versus quantitative explanatory variable
 - ▶ Is the form of the model correct?
 - ▶ Use Wald or LRT tests
 - ▶ Is saturated model better than fitted model?
 - ▶ Deviance GOF test
- ① Are there outliers?
- ▶ Examine standardized residuals: Pearson and Deviance Residuals
 - ▶ Consider other model fit statistics: AIC, BIC
 - ▶ Other issues/concerns in model fitting

Residuals: Pearson and Deviance

- Response (raw) residuals: (*observed – fitted*) proportion

$$\hat{\pi}_{S,i} - \hat{\pi}_{M,i} = \frac{y_i}{m_i} - \hat{\pi}_{M,i}$$

$$\begin{aligned}\hat{\pi}_i &\in (0, 1) \\ \hat{\pi}_i - \hat{\pi}_{M,i} &\in (-1, 1) \\ &\in (-\infty, \infty)\end{aligned}$$

- Standardized residuals:

- (1) Pearson Residuals: uses estimate of s.d. of Y (in denominator)

$$P_{res,i} = \frac{y_i - m_i \hat{\pi}_{M,i}}{\sqrt{m_i \hat{\pi}_{M,i} (1 - \hat{\pi}_{M,i})}}$$

(-2, 2)

- (2) Deviance Residuals: defined so that the sum of the squares of the residuals is the deviance

$$D_{res,i} = \text{sign}(y_i - m_i \hat{\pi}_{M,i})$$

$$\times \sqrt{2 \left\{ y_i \log \left(\frac{y_i}{m_i \hat{\pi}_{M,i}} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i - m_i \hat{\pi}_{M,i}} \right) \right\}}$$

$$\sum_i d_i^2 = \zeta^2$$

= Deviance
of fitted
model.

18/41

Response, Pearson and Deviance Residuals in R

- ▶ Response residuals

```
residuals(fitbl, type="response")
```


(-1,1)

- ▶ Pearson residuals

```
residuals(fitbl, type="pearson")
```

(- ∞ , ∞)

- ▶ Deviance residuals

```
residuals(fitbl, type="deviance")
```

(- ∞ , ∞)

Case Study IV Example: Were there outliers in the data?

	Pearson, $P_{res,i}$	Deviance, $D_{res,i}$
Asymptotic Dist.	$N(0, 1)$	$N(0, 1)$
R code	<code>pearson</code>	<code>deviance</code>
Possible outlier if	$ P_{res,i} > 2$	$ D_{res,i} > 2$
Outlier if	$ P_{res,i} > 3$	$ D_{res,i} > 3$
Under small n	D_{res} closer to $N(0, 1)$ than P_{res}	
$\hat{\pi}$ close to 0 or 1	P_{res} are unstable; related to instability of Wald	

- ▶ Results: Both are $< |2|$, so no outliers

Case IV Residuals

```
rres<-residuals(fitbl, type=c("response"))
pres<-residuals(fitbl, type=c("pearson"))
dres<-residuals(fitbl, type=c("deviance"))
rbind(pis,phats,rres, pres,dres)
```

$$\hat{\pi}_i = \frac{e^{\hat{M}_i}}{1 + e^{\hat{M}_i}}$$

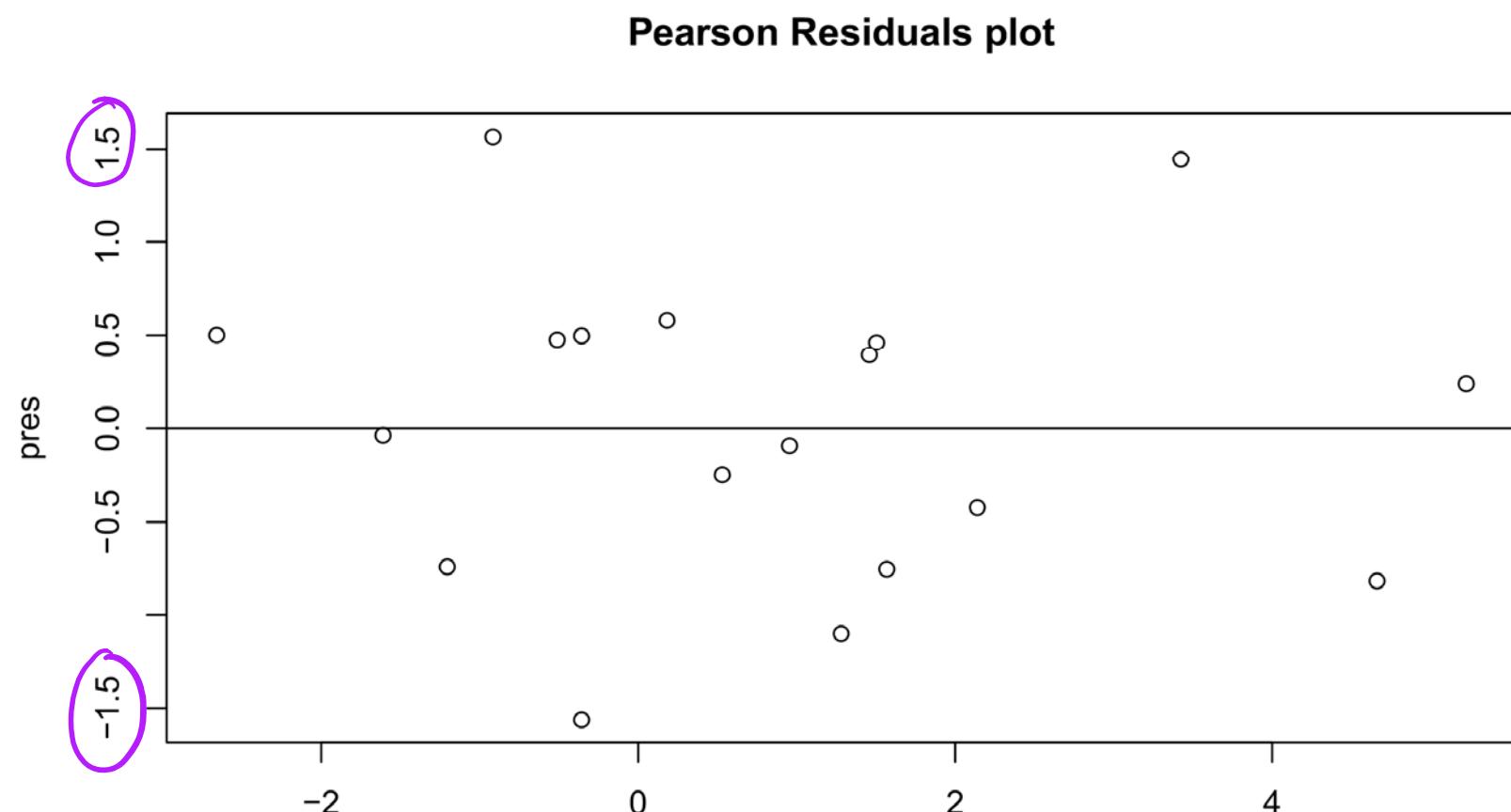
$\hat{\pi}_i$ — $\hat{\pi}_i$

	1	2	3	4	5	6	7	8
## pis	0.066667	0.04478	0.15152	0.11765	0.10714	0.20000	0.18605	0.09677
## phats	0.060173	0.07036	0.09854	0.13800	0.15946	0.16205	0.16389	0.17125
## rres	0.006493	-0.02558	0.05298	-0.02035	-0.05232	0.03795	0.02216	-0.07448
## pres	0.236464	-0.81883	1.44400	-0.42139	-0.75619	0.46058	0.39247	-1.10075
## dres	0.232656	-0.87369	1.34958	-0.43071	-0.79584	0.44746	0.38577	-1.18097
##	9	10	11	12	13	14	15	16
## pis	0.178571	0.18750	0.26667	0.1000	0.29032	0.3125	0.4667	0.24242
## phats	0.185415	0.20524	0.22264	0.2516	0.25158	0.2603	0.2842	0.30185
## rres	-0.006844	-0.01774	0.04403	-0.1516	0.03875	0.0522	0.1825	-0.05943
## pres	-0.093181	-0.24850	0.57969	-1.5622	0.49717	0.4759	1.5673	-0.74367
## dres	-0.093632	-0.25127	0.56727	-1.7173	0.48934	0.4666	1.4954	-0.75939
##	17	18						
## pis	0.325000	0.5000						
## phats	0.327828	0.3998						
## rres	-0.002828	0.1002						
## pres	-0.038101	0.5008						
## dres	-0.038129	0.4957						

s/\sqrt{s}

Case IV Residuals Plot

```
plot(log(Area), pres, main="Pearson Residuals plot")
abline(h=0)
```



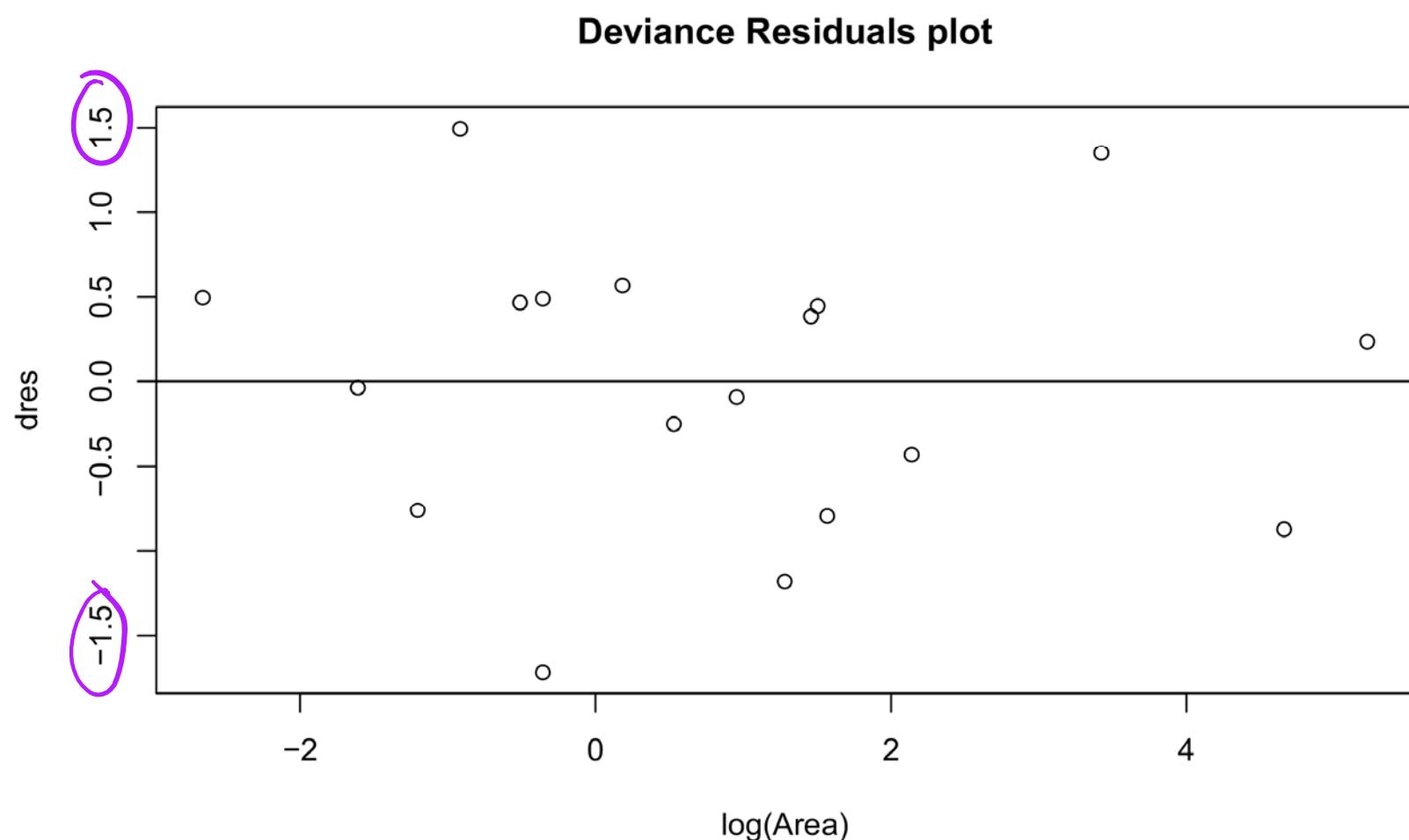
log(Area)

\hat{e}_i

22/41

Case IV Residuals Plot

```
plot(log(Area), dres, main="Deviance Residuals plot")
abline(h=0)
```



Other Model Fit Statistics

- ▶ Useful for comparing models with same response and same data
- ▶ Two popular fit statistics: AIC and BIC; combines log-likelihood with a penalty
 1. Akaike's Information Criterion (AIC)

$$AIC = -2 \log \mathcal{L} + 2(p + 1)$$

2. Schwarz's (Bayesian Information) Criterion (BIC)

$$BIC = -2 \log \mathcal{L} + (p + 1) \log N$$

where

- ▶ p -number of explanatory variables, and
 - ▶ $N = \sum_{i=1}^n m_i$.
- ▶ Example: see AIC, BIC for Case IV model

Case IV Fit Statistics

```
AIC(fitbl)
```

```
## [1] 75.39
```

```
BIC(fitbl)
```

```
## [1] 77.17
```

Problems and Solutions in Logistic Regression

26/41

Logistic Regression Diagnostics

Problems and Complications common to Linear and Logistic Regression

- ▶ *Extrapolation*- don't make inferences/predictions outside range of observed data; model may no longer be appropriate.
- ▶ *Multicollinearity*- highly correlated explanatory variables; difficult to assess individual effects on response. Consequences include:
 - ▶ Unstable fitted equation
 - ▶ Coefficient that should be statistically significant is not
 - ▶ Coefficient may have the wrong sign
 - ▶ Sometimes, large s.e. of $\hat{\beta}$
 - ▶ Sometimes numerical procedure to find MLEs does not converge

Problems and Complications common to Linear and Logistic Regression

- ▶ *Extrapolation*- don't make inferences/predictions outside range of observed data; model may no longer be appropriate.
- ▶ *Multicollinearity*- highly correlated explanatory variables; difficult to assess individual effects on response. Consequences include:
 - ▶ Unstable fitted equation
 - ▶ Coefficient that should be statistically significant is not
 - ▶ Coefficient may have the wrong sign
 - ▶ Sometimes, large s.e. of $\hat{\beta}$
 - ▶ Sometimes numerical procedure to find MLEs does not converge

Problems and Complications common to Linear and Logistic Regression

- ▶ *Influential points*- an observation is influential if its removal substantially changes estimated coefficients (such as, fitted $\hat{\beta}$'s, deviance)
- ▶ *Model Building*- choosing explanatory variables and their forms (eg. polynomial terms, interaction and transformations) tend to overfit the data; should build model on training data and test on test data (cross validation).

Problems and Complications common to Linear and Logistic Regression

- ▶ *Influential points*- an observation is influential if its removal substantially changes estimated coefficients (such as, fitted $\hat{\beta}$'s, deviance)
- ▶ *Model Building*- choosing explanatory variables and their forms (eg. polynomial terms, interaction and transformations) tend to overfit the data; should build model on training data and test on test data (cross validation).

Two problems specific to Logistic Regression

1. Extra-binomial variation

- ▶ variance of Y_i greater than $m_i\pi_i(1 - \pi_i)$
- ▶ also called “over dispersion”
- ▶ does not bias $\hat{\beta}$'s but s.e. of $\hat{\beta}$'s will be too small
(too small p -values, too narrow CIs)

$$Y_i \sim \text{Bin}(m_i, \pi_i)$$

X_1, X_2, \dots, X_{m_i}
 $\stackrel{iid}{\sim} \text{Ber}(\pi_i)$.

$$y = \sum_{i=1}^{m_i} X_i$$

$\hat{\psi}$

$$\hat{\psi} m_i \pi_i (1 - \pi_i)$$

Quasi-binomial

31/41

Example of Extra-binomial variation

- ▶ Suppose X_1, \dots, X_{m_1} are not independent but identical Bernoulli(π)
- ▶ Suppose all pairs (X_i, X_j) have a common correlation ρ
- ▶ Let $Y_1 = \sum_i^{m_1} X_i$

$$\begin{aligned}
 \text{Var}(Y_1) &= \text{Var}\left(\sum_i^{m_1} X_i\right) \\
 &= \sum_i^{m_1} \text{Var}(X_i) + \underbrace{\sum_{i \neq j} \text{Cov}(X_i, X_j)}_{\text{extra variation}} \\
 &= m_1 \pi(1 - \pi) + \sum_{i \neq j} \rho \sqrt{\text{Var}(X_i) \text{Var}(X_j)} \\
 &= m_1 \pi(1 - \pi) + n(n - 1)\rho\pi(1 - \pi), \text{ assume } \rho > 0 \\
 &> m_1 \pi(1 - \pi)
 \end{aligned}$$

Hints of Over dispersion:

- ① Non iid Ber per grp.
- ② Deviations GOF.
(Small p-value).
- ③ Outliers

Estimating Extra-binomial variation

- ▶ Model for variance: $\text{Var}(Y_i) = \psi m_i \pi_i (1 - \pi_i)$
- ▶ Estimate of $\hat{\psi}$: scaled Pearson chi-square statistic,

$$\hat{\psi} = \frac{\sum_i^n P_{\text{res},i}^2}{n - (p + 1)} = \frac{\text{sum of squared Pearson residuals}}{d.f.}$$

$\chi^2_{n-(p+1)}$
 $n-(p+1)$.

- ▶ $\hat{\psi} >> 1$ indicates evidence of overdispersion
- ▶ ψ does not affect $E(Y_i)$, hence using overdispersion does not change $\hat{\beta}$
- ▶ $SE(\hat{\beta})$ is multiplied by $\sqrt{\hat{\psi}}$,

$$\underline{SE_\psi(\hat{\beta})} = \sqrt{\hat{\psi}} \underline{SE_{\psi=1}(\hat{\beta})}$$

- ▶ Overdispersion does not apply to Bernoulli data. If y_i only takes on 0 or 1, then it must be $\text{Bernoulli}(\pi_i)$ and its variance must be $\pi_i(1 - \pi_i)$ (McCullagh and Nelder(1989)).

Case Study IV: Logistic Model with logged explanatory variable

```
fitbl<-glm(cbind(Extinct,NExtinct)~log(Area), family=binomial, data=krunnit)
summary(fitbl)
```

```
##  
## Call:  
## glm(formula = cbind(Extinct, NExtinct) ~ log(Area), family = binomial,  
##       data = krunnit)  
##  
## Deviance Residuals:  
##      Min        1Q    Median        3Q       Max  
## -1.71726  -0.67722   0.09726   0.48365   1.49545  
##  
## Coefficients:  
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.19620   0.11845 -10.099 < 2e-16 ***  
## log(Area)   -0.29710   0.05485  -5.416 6.08e-08 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 45.338 on 17 degrees of freedom  
## Residual deviance: 12.062 on 16 degrees of freedom  
## AIC: 75.394  
##  
## Number of Fisher Scoring iterations: 4
```

$$\text{Var}(Y_i) = \pi_i \pi_i (1-\pi_i)$$

Case IV Estimating ψ

```
(psihat=sum(residuals(fitbl, type="pearson")^2/fitbl$df.residual))  
  
## [1] 0.7326  
  
summary(fitbl, dispersion=psihat)  
  
##  
## Call:  
## glm(formula = cbind(Extinct, NExtinct) ~ log(Area), family = binomial,  
##       data = krunnit)  
##  
## Deviance Residuals:  
##      Min        1Q    Median        3Q       Max  
## -1.7173  -0.6772   0.0973   0.4837   1.4954  
##  
## Coefficients:  
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.1962     0.1014 -11.80 < 2e-16 ***  
## log(Area)   -0.2971     0.0469  -6.33 2.5e-10 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 0.7326)  
##  
## Null deviance: 45.338 on 17 degrees of freedom  
## Residual deviance: 12.062 on 16 degrees of freedom
```

$$\hat{\psi} \text{SE}(\hat{\beta}) = S_{\hat{\beta}}(\hat{\beta})$$

Case IV: As a Quasi-Binomial

```
fitbl2=glm(cbind(Extinct,NExtinct)~log(Area), family=quasibinomial)
summary(fitbl2)
```

```
##  
## Call:  
## glm(formula = cbind(Extinct, NExtinct) ~ log(Area), family = quasibinomial)  
##  
## Deviance Residuals:  
##      Min        1Q    Median        3Q       Max  
## -1.7173   -0.6772    0.0973    0.4837    1.4954  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.1962    0.1014 -11.80  2.6e-09 ***  
## log(Area)   -0.2971    0.0469  -6.33  1.0e-05 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for quasibinomial family taken to be 0.7326)  
##  
## Null deviance: 45.338 on 17 degrees of freedom  
## Residual deviance: 12.062 on 16 degrees of freedom  
## AIC: NA  
##  
## Number of Fisher Scoring iterations: 4
```

$$\text{Var}(y_i) = \psi m_i \pi_i (1-\pi_i).$$

Same as p-35

Two problems specific to logistic regression

2. Complete and Quasi-complete separation

- ▶ *Complete separation:*

- ▶ one or a linear combination of explanatory variables perfectly predict whether $Y = 1$ or $Y = 0$
- ▶ In Binary response, when $y_i = 1$, $\hat{y}_i = 1$, then $\sum_{i=1}^n \{y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\} = 0$.
- ▶ MLE's cannot be computed

- ▶ *Quasi-complete separation:*

- ▶ explanatory variables predict $Y = 1$ or $Y = 0$ almost perfectly (just a few points wrong)
- ▶ MLE's are numerically unstable

SOLUTION: simplify the model. Other options- penalized maximum likelihood, exact logistic regression, bayesian methods

Using Logistic Regression for Classification

38/41

Logistic Regression Diagnostics

Using Logistic Regression for Classification

- **Want:** predict outcome as

$$y^* | (x_1^*, x_2^*, \dots, x_p^*) = \begin{cases} 1 \\ 0 \end{cases}$$

- **Do:** calculate $\hat{\pi}_M^*$ - the estimated probability that $y^* = 1$ based on the fitted model given $X_1 = x_1^*, X_2 = x_2^*, \dots, X_p = x_p^*$.
From this we want to predict that

$$y^* = \begin{cases} 1 & \text{if } \hat{\pi}_M^* \text{ is large} \\ 0 & \text{if } \hat{\pi}_M^* \text{ is small} \end{cases}$$

- **Need:** choose a cut-off probability to distinguish between large and small.

Classification: Approaches to choosing a threshold

Approach 1 - Set cut-off probability as 0.5

- ▶ If $\hat{\pi}_M^* > 0.5$, classify y^* as 1
- ▶ Useful if there are equal numbers of 1's and 0's
- ▶ Useful if false negatives and false positives are equally bad.

Classification: Approaches to choosing a threshold

Approach 2- Find “best” cut-off probability from data.

- ▶ Try different cut-offs and see which gives fewest incorrect classifications
- ▶ Useful if proportions of 1's and 0's in data reflect their relative proportions in the population
- ▶ Likely to overestimate the proportions of correct predictions that model makes. Then, one should assess model correct classification rates on different data than was used to fit the model.

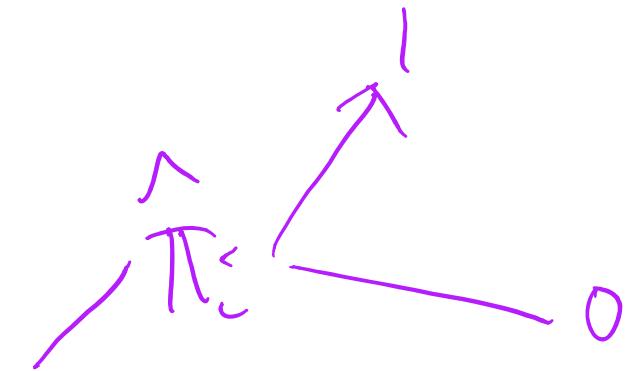
STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee



February 27, 2018



Using Logistic Regression for Classification

Classification and Review

Using Logistic Regression for Classification

- **Want:** predict outcome as

$$y^* | (x_1^*, x_2^*, \dots, x_p^*) = \begin{cases} 1 \\ 0 \end{cases}$$

- **Do:** calculate $\hat{\pi}_M^*$ - the estimated probability that $y^* = 1$ based on the fitted model given $X_1 = x_1^*, X_2 = x_2^*, \dots, X_p = x_p^*$.
From this we want to predict that

$$y^* = \begin{cases} 1 & \text{if } \hat{\pi}_M^* \text{ is large} \\ 0 & \text{if } \hat{\pi}_M^* \text{ is small} \end{cases}$$

- **Need:** choose a cut-off probability to distinguish between large and small.

Classification: Approaches to choosing a threshold

Approach 1 - Set cut-off probability as 0.5

- ▶ If $\hat{\pi}_M^* > 0.5$, classify y^* as 1
- ▶ Useful if there are equal numbers of 1's and 0's
- ▶ Useful if false negatives and false positives are equally bad.

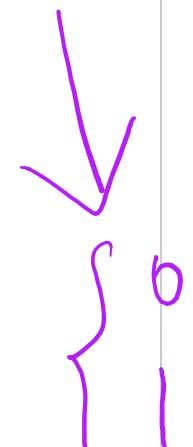
Classification: Approaches to choosing a threshold

Approach 2- Find “best” cut-off probability from data.

- ▶ Try different cut-offs and see which gives fewest incorrect classifications
- ▶ Useful if proportions of 1's and 0's in data reflect their relative proportions in the population
- ▶ Likely to overestimate the proportions of correct predictions that model makes. Then, one should assess model correct classification rates on different data than was used to fit the model.

More reliable way to find a cut-off prob. is to use cross-validation.

- train ← estimates
- test ← test estimates

$$0 < \hat{\pi} < 1$$


Confusion Matrix

Prediction	Truth		Prop (row)
	Positive ($Y = 1$)	Negative ($Y = 0$)	
Positive	TP	FP	$PPV = \frac{TP}{TP+FP}$
Negative	FN	TN	$NPV = \frac{TN}{TN+FN}$
Prop (column)	Sensitivity= $TPR = \frac{TP}{TP+FN}$	Specificity= $TNR = \frac{TN}{TN+FP}$	

Estimate

Observed
type I error

- ▶ TP: true positive; TN: true negative
- ▶ FP: false positive (type I error); FN: false negative (type II error)
- ▶ PPV: precision or positive predictive value; false discovery rate=1-PPV
- ▶ NPV: negative predictive value; false omission rate=1-NPV

1. Sensitivity (True Positive Rate, TPR)- hit rate
2. Specificity (True Negative Rate, TNR)- prop. of correctly classified negatives
3. False Positive Rate, FPR=1-TNR, fall-out rate
4. False Negative Rate, FNR=1-TPR, miss rate
5. Classification rate=($TN+TP$)/($TP+FN+FP+TN$)); accuracy

Diagnostic Accuracy



Choose a cut-off probability based on one of the 5 criteria for success of classification that is most important to you.

- ▶ High Sensitivity (TPR) makes good screening test.
- ▶ High Specificity (TNR) makes a good confirmatory test.
- ▶ A screening test followed by a confirmatory test is a good (but expensive) diagnostic procedure.

Confusion Matrix

► From Wikipedia

		True condition				
		Total population	Condition positive	Condition negative	Prevalence = $\frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$
Predicted condition	Predicted condition positive	True positive, Power	False positive, Type I error		Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative		False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$		Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$	F ₁ score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
	False negative rate (FNR), Miss rate = $\frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$		Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$		

https://en.wikipedia.org/wiki/Confusion_matrix

Classification and Review

Least Squares Regression vs Logistic Regression

1-way & 2-way ANOVA

$$\epsilon_i \sim N(0, \sigma^2)$$

	(Ordinary) Least Squares	(Binomial) Logistic
Response, Y	Normal	# of successes in m trials $\sim \text{Bin}(m, \pi)$.
Variance	Equal for each level of X	$mp(1 - p)$ for each level of X $M_i\pi_i(1 - \pi_i)$.
Model	$\mu_y = \beta_0 + \beta_1 X$	$\log(\frac{\mu}{1-\mu}) = \beta_0 + \beta_1 X$
Model fitting	Least Squares	MLE
Exploratory plot	X vs Y (add line)	logit vs \underline{X}
Comparing models	Partial F-test AIC/BIC Residuals	LRT/Deviance tests AIC/ BIC (Pearson, Deviance) Residuals Wald. χ^2
Interpreting	β_1 : change in μ_y for unit change in X	e^{β_1} : % change in odds for unit change in X

Bonferroni vs Tukey's

Tuesday, February 27, 2018

10:48 AM

Bonf:

$$(\bar{y}_i - \bar{y}_j) \pm t_{\alpha^*/2, df \text{ Error}}$$

$$S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad \alpha^* = \frac{\alpha}{k}$$

Critical value
from Tukey's Range distribution

Tukey's:

$$(\bar{y}_i - \bar{y}_j) \pm q_{df \text{ Error}, k}^{**}$$

$$S \sqrt{\frac{1}{h_i} + \frac{1}{h_j}} \quad k = \binom{G}{2}$$

(all possible pairs.)

$$S = \sqrt{MSE_{\hat{F}}} = \hat{F} = \sqrt{\hat{f}^2}$$

(Residual standard error in R)

How does the estimated logistic model change as we change the "success" and reference level of X ?

$$\pi_s = P(\text{success})$$

$$\text{logit}(\hat{\pi}_s) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2, \text{ref} = 1$$

Ⓐ Changing "success"

$$\pi_f = P(\text{failure})$$

$$\text{logit}(\hat{\pi}_f) = \frac{1}{\text{logit}(\hat{\pi}_s)} = \frac{1}{\log\left(\frac{\hat{\pi}_s}{\hat{\pi}_f}\right)} = -\log\left(\frac{\hat{\pi}_s}{\hat{\pi}_f}\right) = -\hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2$$

$$\text{logit}(\hat{\pi}_s) = \log\left(\frac{\hat{\pi}_s}{1 - \hat{\pi}_f}\right) = \frac{\log \hat{\pi}_s}{\log \frac{\hat{\pi}_s}{\hat{\pi}_f}} = \log \hat{\pi}_s - \log \hat{\pi}_f$$

③ Changing ref. level of X

$$\text{logit}(\hat{\pi}_s) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_{2,\text{ref}1}$$

ref 1 = 1

$$\begin{cases} \text{logit}(\hat{\pi}_s) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X_1 & (\text{ref } 1 = 1) \\ \text{logit}(\hat{\pi}_s) = \hat{\beta}_0 + \hat{\beta}_1 X_1 & (\text{ref } 1 = 0) \end{cases}$$

$$\text{logit}(\hat{\pi}_s) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 1_{\text{ref}1}$$

See Case III - Donner Party Eq.

$$\text{logit}(\hat{\pi}_s) = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X_1 - \hat{\beta}_2 1_{\text{ref}0}$$