STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

January 23-25, 2018

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: $\epsilon \sim N(0, \sigma^2 \mathbf{I})$
- Review: One-way ANOVA
  - Special case of a GLM
  - One-way classification/ One factor with $G \geq 2$ levels

- Multiple Comp.

- What if we have more than one factor?
  - Main and Interaction effect of factors on Y?
  - Assumptions?
  - Visualizations?
  - Analyses?

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 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.

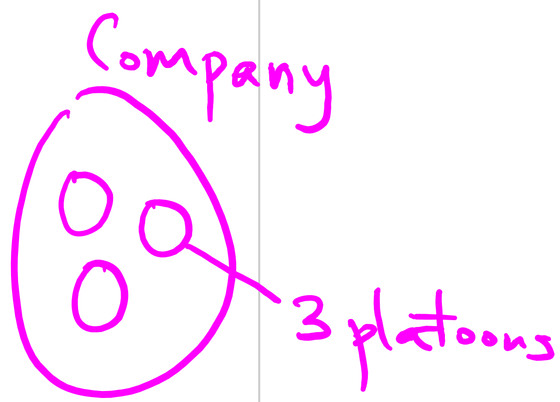
Case Study II-The Pygmalion Effect

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

Company		Treatments 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

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)

Case Study II Variables

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

where

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

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

Case Study II Linear Model

Full Model:

1st platoon, PYG, Comp 1

$$
\begin{aligned}
Y_i = \beta_0 + \beta_1 \overset{1}{\mathbb{1}_{PYG,i}} + \beta_2 \overset{1}{\mathbb{1}_{COMP_1,i}} \\
+ \beta_3 \mathbb{1}_{COMP_2,i} \; 0 \\
+ \cdots \\
+ \beta_{10} \mathbb{1}_{COMP_9,i} \; 0
\end{aligned}
$$

$$
\boxed{
\begin{aligned}
+ \beta_{11} \overset{1}{\mathbb{1}_{PYG,i}} \times \overset{1}{\mathbb{1}_{COMP_1,i}} \\
+ \beta_{12} \mathbb{1}_{PYG,i} \times \mathbb{1}_{COMP_2,i} \; 0 \\
+ \cdots \\
+ \beta_{19} \mathbb{1}_{PYG,i} \times \mathbb{1}_{COMP_9,i} \; 0
\end{aligned}
}
$$

$$+\epsilon_i$$

$$E[y_1 \mid (\text{treat}_1, \text{company}_1)] = \beta_0 + \beta_1 + \beta_2 + \beta_{11}$$

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

(Pyg - Control)

Company	Pygmalion($\mathbb{1}_{PYG,i} = 1$)	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	$\underbrace{\beta_0 + \beta_1} + \beta_3 + \beta_{12}$	$\beta_0 + \beta_3$	$\beta_1 + \beta_{12}$
3			
4			
5			
6			
7			
8			
9	$\beta_0 + \beta_1 + \beta_{10} + \beta_{19}$	$\beta_0 + \beta_{10}$	$\beta_1 + \beta_{19}$
10	$\underbrace{\beta_0 + \beta_1}$	β_0	β_1

Question 1: Does mean treatment effect differ with *Company*?

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

Alternative Hypothesis, H_a : at least 1 β is not 0

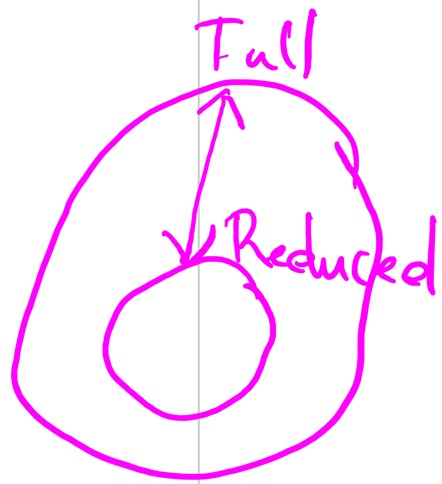
Overall versus Partial F-tests in Two-way ANOVA

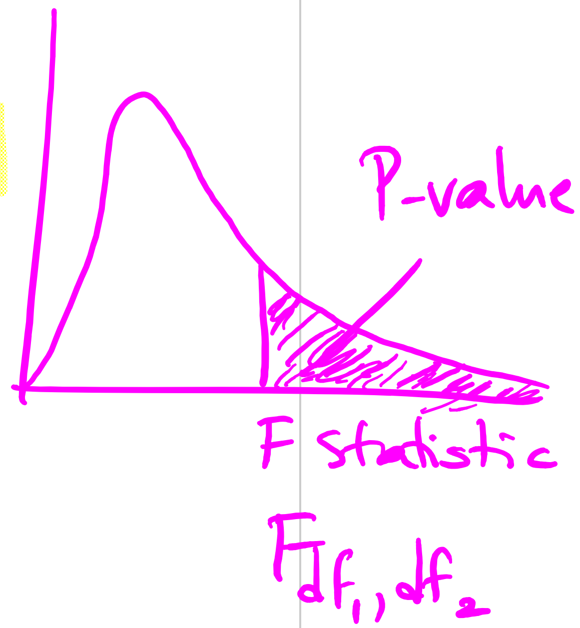
all

- Full ▸ Overall test: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{dfMODEL} = 0$
- Reduced ▸ Partial test: H_0 : a subset of $\{\beta_1, \beta_2, \cdots, \beta_{dfMODEL}\} = 0$
- Approach: Fit full model (with all explanatory variables) and reduced (without variables whose coefficients you are testing) model
  - Test statistic:

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

  - If H_0 is true, F is an observation from F distribution with $df = (\#$ of β's being tested, dfERROR of full model)





Case Study II: Testing interaction

- FULL:

19

$$\text{full=lm(score}\sim\text{company*treat)}$$

- Reduced:

10

$$\text{reduced=lm(score}\sim\text{company+treat)}$$

$19 - 10 = 9$

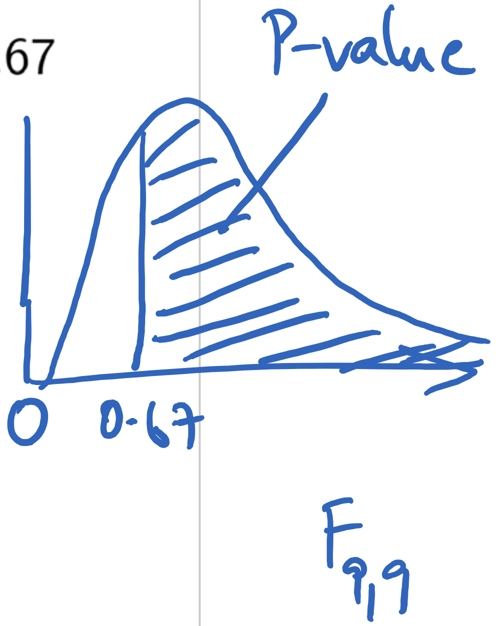
$\beta_{11}, \beta_{12}, \cdots, \beta_{19}$

- Partial F-test (Refer to R output)
  - Test statistic:

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

  - 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.

$\Rightarrow$ Fit Additive model



- No evidence that treatment effect differs with *Company*.

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:

$$\begin{aligned} Y_i = \beta_0 + \beta_1 \mathbb{1}_{PYG,i} &+ \beta_2 \mathbb{1}_{COMP_1,i} \\ &+ \beta_3 \mathbb{1}_{COMP_2,i} \\ &+ \cdots \\ &+ \beta_{10} \mathbb{1}_{COMP_9,i} + \epsilon_i \end{aligned}$$

Treat

Company

Case Study II: Additive Model Expected Response

Company	Pygmalion($\mathbb{1}_{PYG,i} = 1$)	Control($\mathbb{1}_{PYG,i} = 0$)	Treatment effect
	Treatment		
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?

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

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
```

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
```

Case Study II: Additive Model-Testing main effects

	Test 1	Test 2
Null	$H_0 : \beta_1 = 0$	$H_0 : \beta_2 = \beta_3 = \cdots = \beta_{10} = 0$
Alt	$H_a : \beta_1 \neq 0$	H_a :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)
```

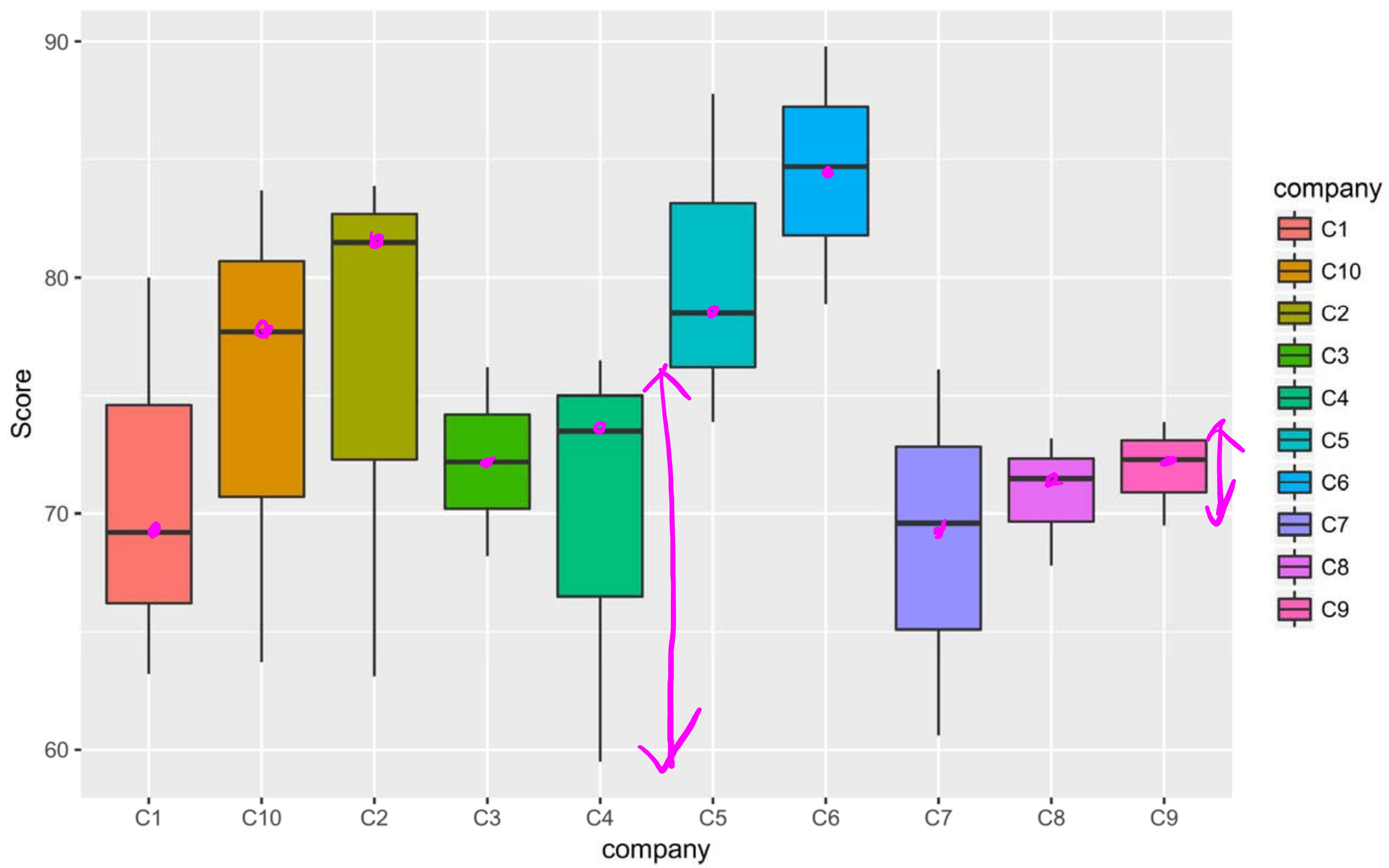
```
##   Company      Treat Score
## 1      C1 Pygmalion  80.0
## 2      C1   Control  63.2
## 3      C1   Control  69.2
## 4      C2 Pygmalion  83.9
## 5      C2   Control  63.1
## 6      C2   Control  81.5
```

29

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

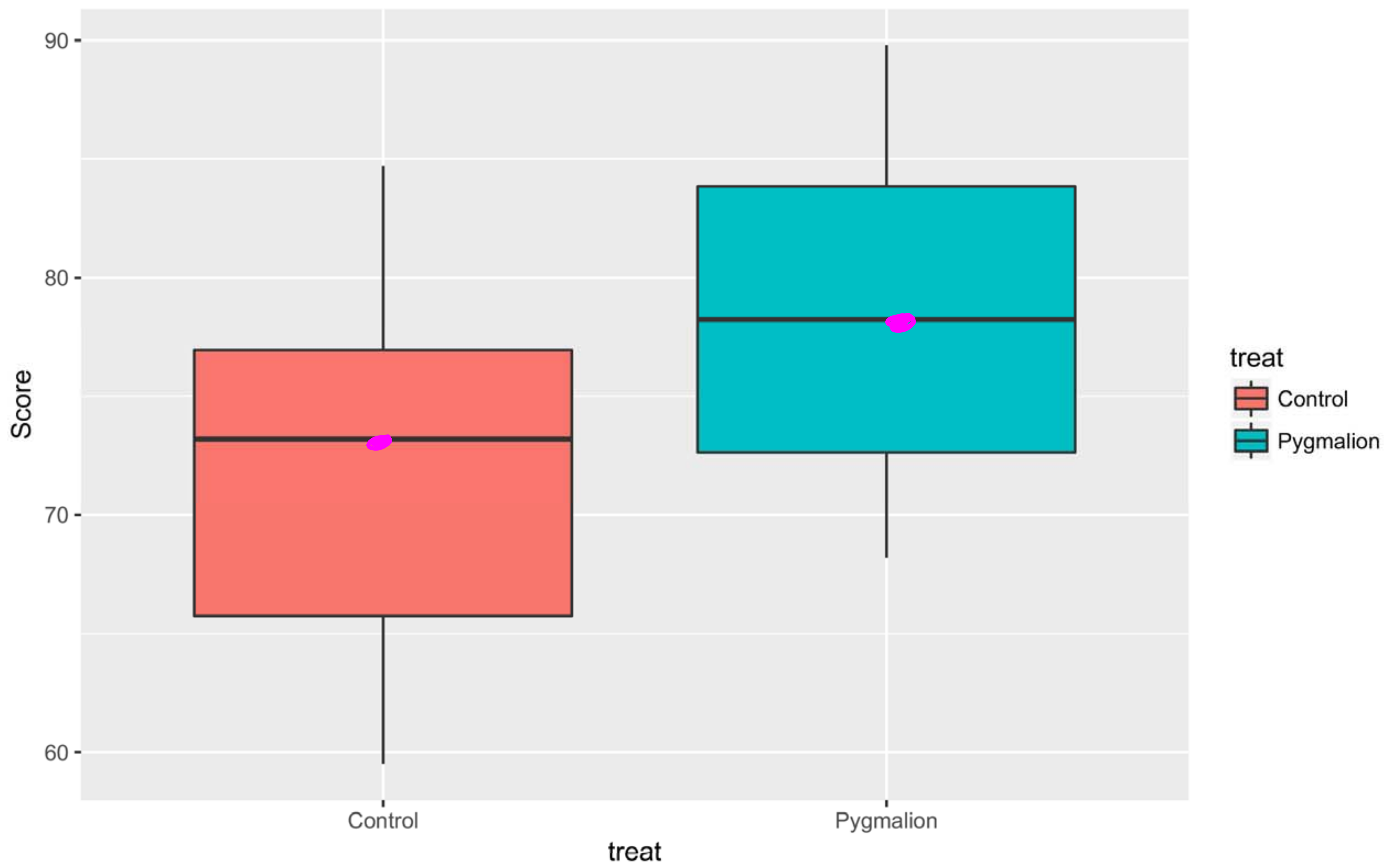
Case Study 2: Visualizing the data

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



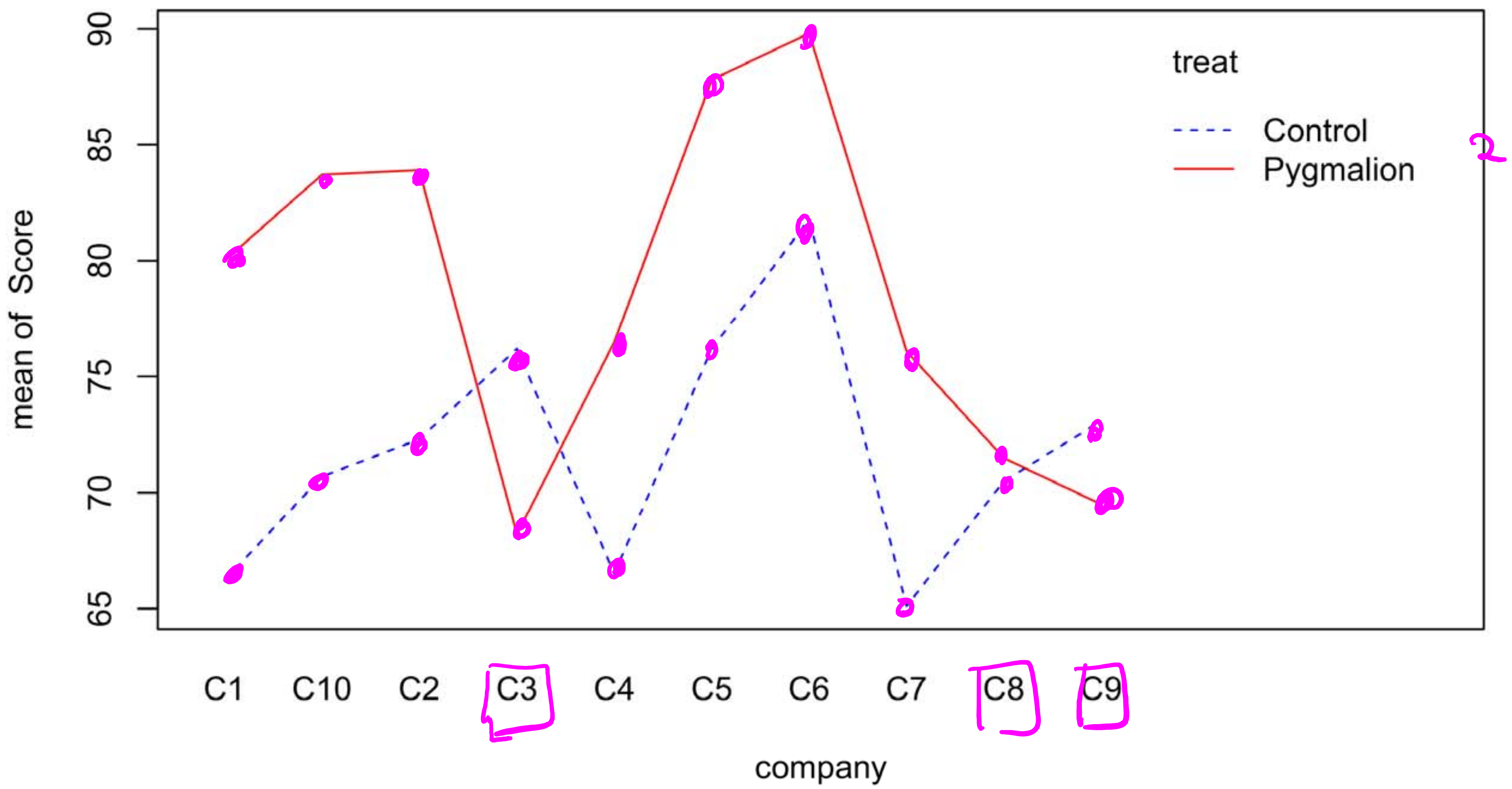
Case Study 2: Visualizing the data

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



Case Study 2: Interaction plots

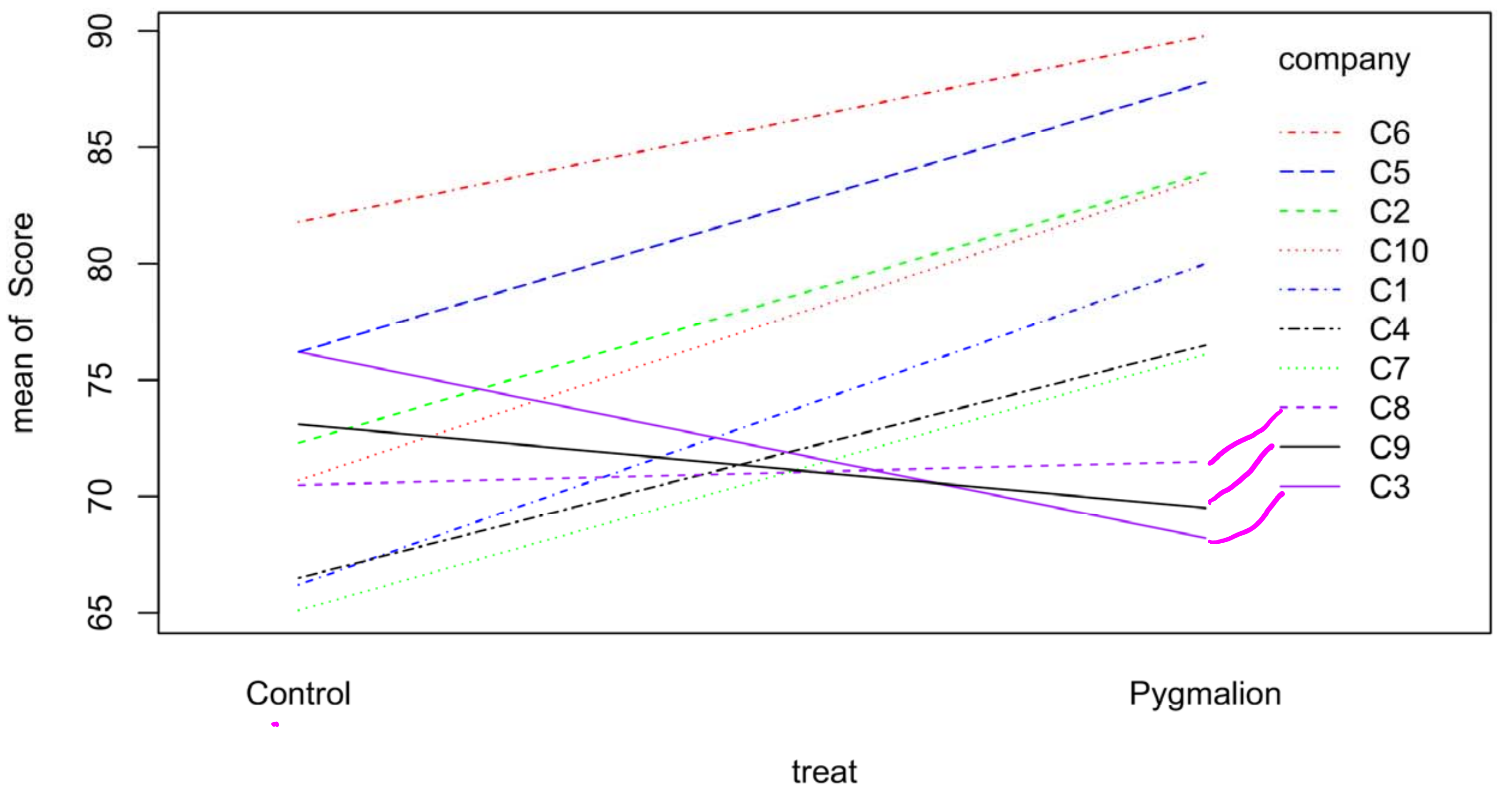
```
interaction.plot(company,treat,Score,col=c("blue","red"))
```



2x10 = 20

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

```r
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

```r
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
```

```r
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           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

10-1

2-1 =

$SSReg_{FULL}$

RSS_{FULL}

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

df

$G_1 - 1$

$G_2 - 1$

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

$N - G_1 G_2$

$N-1$

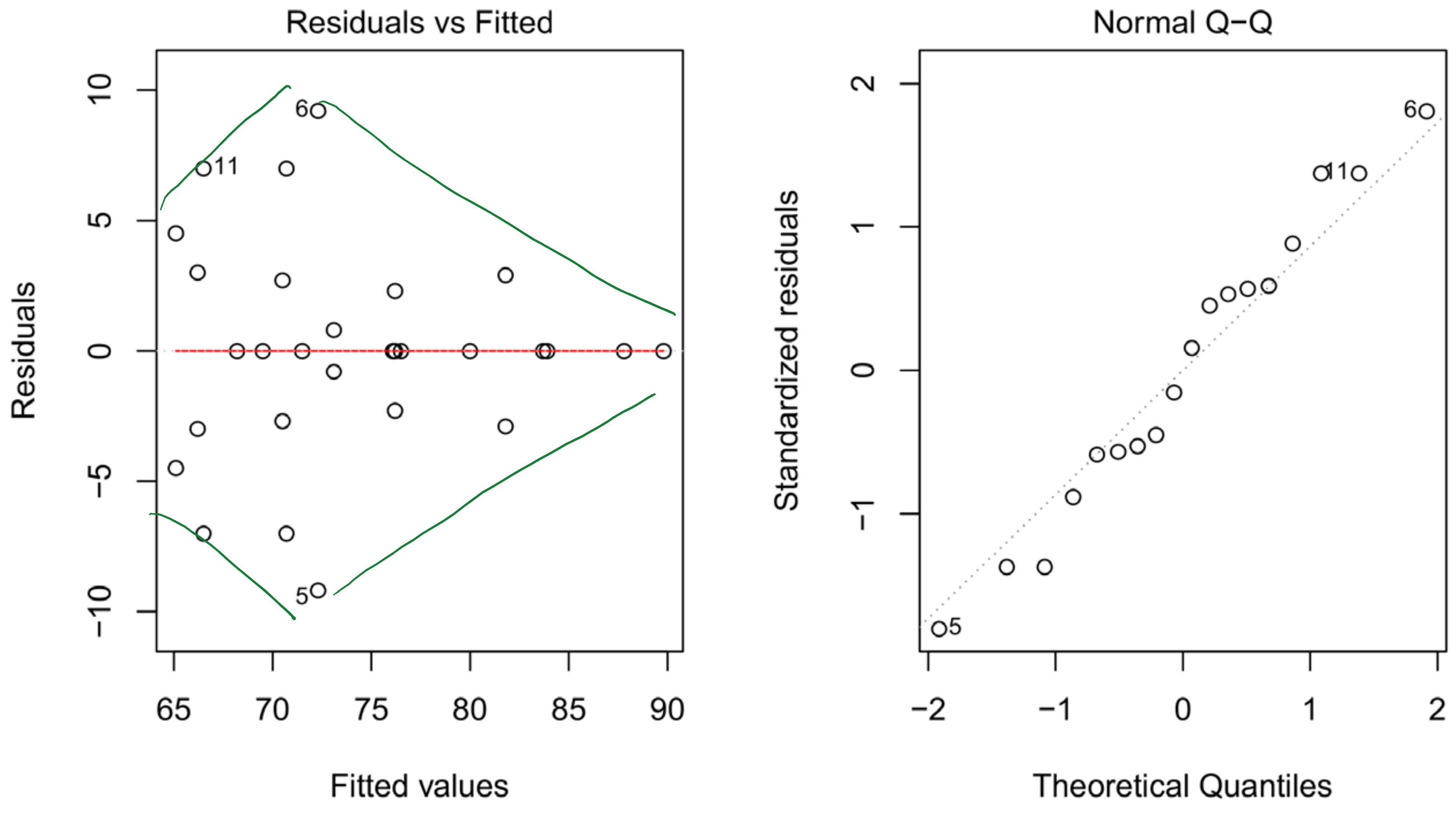
DF Total = N-1 = # of observations - 1 = 29-1 = 28

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

```r
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))
```

Red: Score ~ company

Full: Score ~ comp + 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

$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$

$F = 7.8354$

p-value $= 0.012$

Case Study 2: Additive model summary

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

Red: Score ~ treat

Full: 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
##
```

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
```

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

$F = 1.75$

$p = 0.148$ (large)

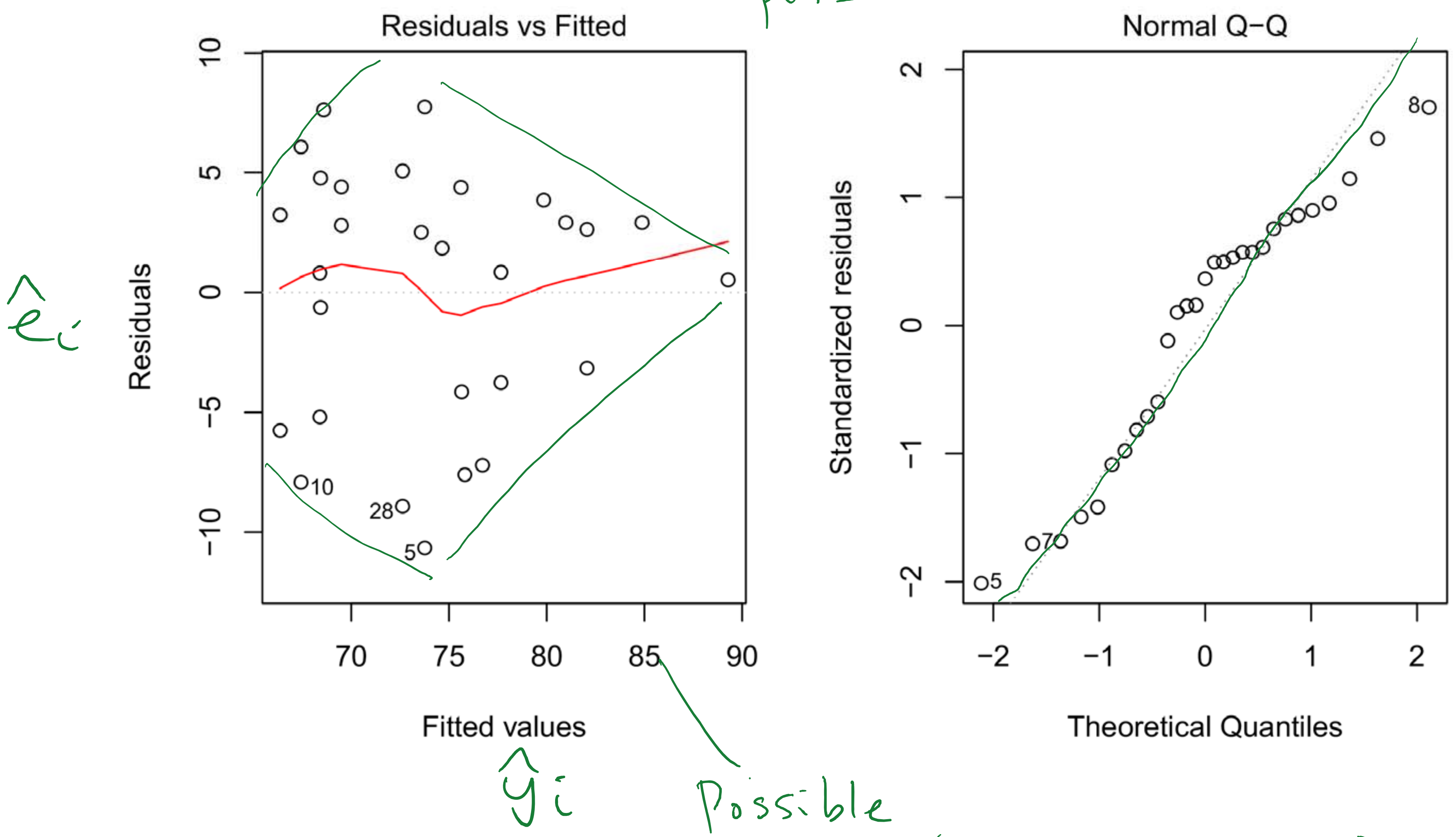
Evidence of no diff in companies.

Case Study 2: Checking assumptions

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

4 plots

1st 2 plots



Normality satisfied

Possible decreasing variance $\Rightarrow$ Weighted L.S.