MATH411 | Fall 2018 | Homework 4 (Due: Monday in class, 11/26/2018)

Paul Tomosky

# Problem 1 (MLR)

So-called “Funds of Hedge Funds” (FoHF), i.e. portfolios of hedge funds, have different investment strategies with specific returns and risk properties. When such a product is evaluated it is important for the investor to choose the investment style that fits his needs.

One approach to assess the investment strategy of a FoHF as an outsider is to perform a style analysis based on the returns. Using a regression model (also called multi-factor model in the financial industry) one aims to explain the returns of the FoHF with the returns of the so-called subindices of hedge funds (Long Short Equity, Fixed Income Arbitrage, Global Macro, etc.). The estimated parameters are indications for the chosen investment strategy. Note that not all investment strategies are present due to the construction of FoHFs.

The file FoHF.rda contains the monthly returns of one FoHF and the hedge fund subindices of EDHEC from January 1997 until December 2004. The meaning of the individual predictors is as follows:

* RV Relative value
* CA Convertible Arbitrage
* FIA Fixed Income Arbitrage
* EMN Equity Market Neutral
* ED Event Driven Multistrategy
* DS Distressed Securities
* MA Merger Arbitrage
* LSE Long Short Equity
* GM Global Macro
* EM Emerging Markets
* CTA CTA / Managed Futures
* SS Short Selling

FoHF = load(file.choose())

FoHF %>% View()

1. **Fit a model containing all predictors. Look at the output of summary. What conclusion can you draw with respect to the investment strategy of this FoHF when you consider the estimated coefficients, the p-values, the global F-test and the multiple R-squared?**

FoHF\_fit\_1 = lm(FoHF ~ ., data = FoHF) # [RV, CA, FIA, CTA] significant

summary(FoHF\_fit\_1)

> summary(FoHF\_fit\_1)

Call:

lm(formula = FoHF ~ ., data = FoHF)

Residuals:

Min 1Q Median 3Q Max

-0.0185186 -0.0031189 0.0004069 0.0035469 0.0148925

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.002256 0.001299 -1.736 0.0862 .

RV -0.388854 0.171151 -2.272 0.0257 \*

CA 0.238653 0.104522 2.283 0.0250 \*

FIA 0.363010 0.087832 4.133 8.51e-05 \*\*\*

EMN 0.184766 0.197475 0.936 0.3522

ED 0.314914 0.215792 1.459 0.1482

DS -0.007699 0.124324 -0.062 0.9508

MA -0.028413 0.169406 -0.168 0.8672

LSE 0.153636 0.099548 1.543 0.1266

GM 0.127093 0.086897 1.463 0.1474

EM 0.049183 0.035065 1.403 0.1645

CTA 0.159225 0.037304 4.268 5.20e-05 \*\*\*

SS 0.032630 0.023424 1.393 0.1673

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.006563 on 83 degrees of freedom

Multiple R-squared: 0.8076, Adjusted R-squared: 0.7798

F-statistic: 29.03 on 12 and 83 DF, p-value: < 2.2e-16

RV: for each increase by 1 of FOFH, RV decreases by 0.388 if no other predictors change

CA: for each increase by 1 of FOFH, CA increases by 0.238 if no other predictors change

FIA: for each increase by 1 of FOFH, FIA increases by 0.363 if no other predictors change

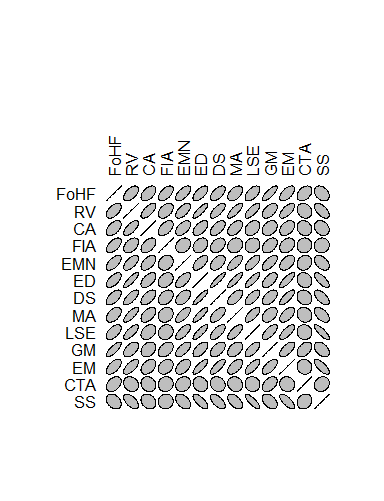
CTA: for each increase by 1 of FOFH, CTA increases by 0.159 if no other predictors change

The f-statistic p-value is significantly small so we can reject the null hypothesis and we can accept that this model is significant.

adjusted r squared = 0.8, is close to one but not quite explain all of the variance

1. [1] Check whether this model is valid or whether any assumptions are violated. [2] Also test whether there are problems with respect to multicollinearity and [3] whether all predictors have been entered into the model in the correct form.

plotcorr(cor(FoHF))

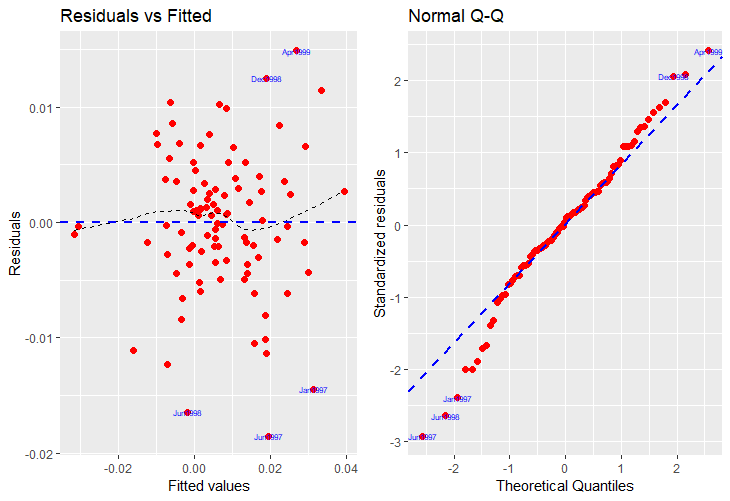


autoplot(FoHF\_fit\_1, which = 1:2, colour = 'red', size = 2,

smooth.colour = 'black', smooth.linetype = 'dashed',

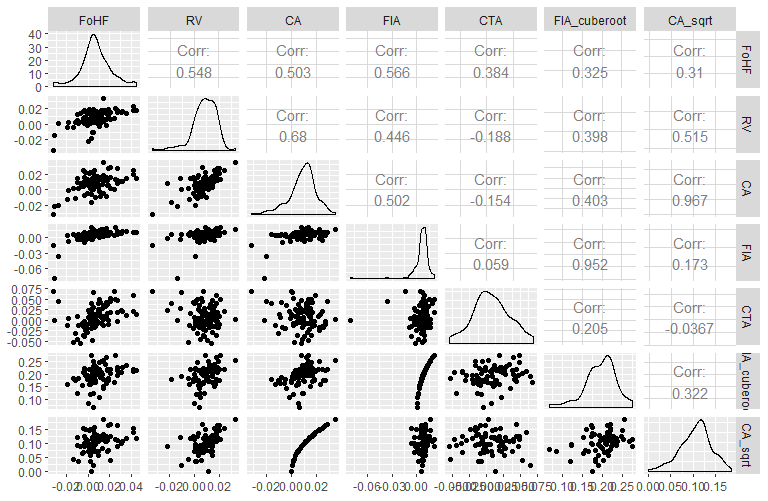
ad.colour = 'blue', ad.size = 1,

label.size = 2, label.n = 5, label.colour = 'blue')



1. If you have solved the previous subproblem correctly, you will have found some issues. Formulate a strategy how those can be fixed in order to obtain a valid and interpretable result.

the residual plot is randomly scattered the points are evenly scattered and are equally distributed. according the the Q-Q plot is light tailed and skewed slightly to the right. This will need to be fixed in order for this model to work.



I transformed FIA with cube root and CA with squareroot

> FOHF\_fit\_4 = lm(FoHF ~ RV + CA + CA\_sqrt + FIA + FIA\_cuberoot + CTA, data = FoHF)

> summary(FOHF\_fit\_4)

Call:

lm(formula = FoHF ~ RV + CA + CA\_sqrt + FIA + FIA\_cuberoot +

CTA, data = FoHF)

Residuals:

Min 1Q Median 3Q Max

-0.0173030 -0.0048539 -0.0008719 0.0023723 0.0255014

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.01738 0.01105 -1.572 0.1209

RV 0.70783 0.16277 4.349 5.02e-05 \*\*\*

CA -0.70673 0.52324 -1.351 0.1816

CA\_sqrt 0.18402 0.10393 1.771 0.0814 .

FIA -0.88768 0.71400 -1.243 0.2183

FIA\_cuberoot 0.07690 0.07812 0.984 0.3287

CTA 0.29335 0.04033 7.274 6.13e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.00794 on 64 degrees of freedom

(25 observations deleted due to missingness)

Multiple R-squared: 0.6044, Adjusted R-squared: 0.5673

F-statistic: 16.29 on 6 and 64 DF, p-value: 2.77e-11

1. Perform variable selection using the BIC criterion.

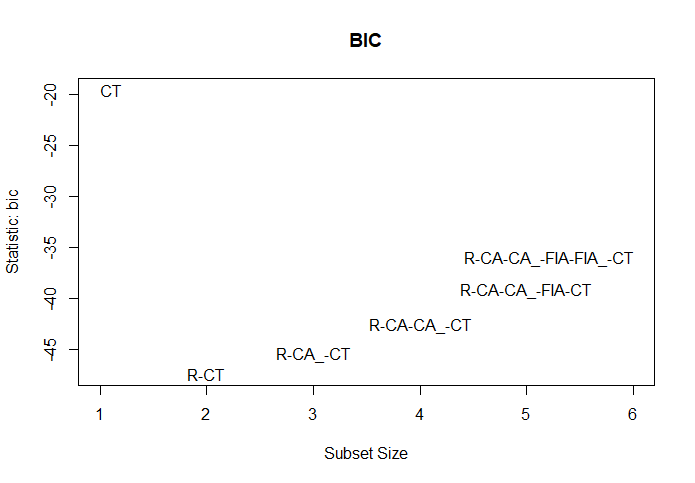
bestsubs = regsubsets(FoHF ~ RV + CA + CA\_sqrt + FIA + FIA\_cuberoot + CTA,

data = FoHF,

nbest = 1, # 1 best model for each number of predictors

method = "exhaustive")

res.legend = subsets(bestsubs, statistic = "bic", legend = FALSE, main = "BIC")



# Problem 2 (MLR & CV)

The file CustomerWinBack.rda provides a dataframe called cwb. It contains information about how long could a company hold costumers that cancelled the contract at some point in the past and re-opened their contracts afterwards. There are 295 observations of the following variables:

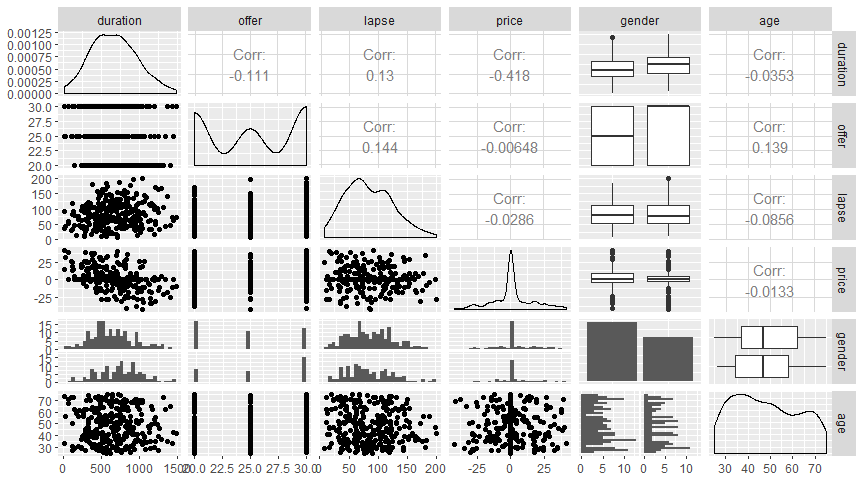
* duration target variable, duration of the customer relationship in days
* offer value of the present offered at re-acquisition
* lapse time until the customer could be re-acquired
* price offered price change in comparison to the first contract
* gender gender. 0 female and 1 male
* age age of the customer

The goal is to find a good model for the duration of the new customer relationship. Since we are primarily interested in an accurate prediction, we use cross validation to evaluate the predictive performance.

1. Change the variable gender to a factor variable and re-label its levels using “Female” and “Male”.

cwb$gender = as.factor(cwb$gender)

1. Make a scatter plot matrix of this data. Comment on any necessary patterns, variable transformations that are needed.



To my understanding, looking at the summary output from summary(cwb) I can see that all the means and medians are very close to each other…

> summary(cwb)

duration offer lapse price gender

Min. : 16.0 Min. :20.0 Min. : 8.00 Min. :-41.760 0:169

1st Qu.: 454.5 1st Qu.:20.0 1st Qu.: 53.00 1st Qu.: -3.530 1:126

Median : 655.0 Median :25.0 Median : 77.00 Median : 0.000

Mean : 668.8 Mean :25.2 Mean : 84.31 Mean : 1.295

3rd Qu.: 870.5 3rd Qu.:30.0 3rd Qu.:112.00 3rd Qu.: 6.215

Max. :1477.0 Max. :30.0 Max. :199.00 Max. : 41.040

age

Min. :25.00

1st Qu.:36.00

Median :47.00

Mean :48.09

3rd Qu.:61.00

Max. :75.00

1. Fit the following models:

* fit1 OLS with all variables.

cwb\_fit\_1 = lm(duration ~ ., data = cwb)

> summary(cwb\_fit\_1)

Call:

lm(formula = duration ~ ., data = cwb)

Residuals:

Min 1Q Median 3Q Max

-538.14 -196.27 -34.79 182.57 744.55

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 836.15414 101.51985 8.236 6.21e-15 \*\*\*

offer -11.91067 3.70155 -3.218 0.00144 \*\*

lapse 1.09216 0.38843 2.812 0.00527 \*\*

price -8.32047 1.03278 -8.056 2.08e-14 \*\*\*

gender1 113.45371 31.73589 3.575 0.00041 \*\*\*

age 0.06461 1.08476 0.060 0.95255

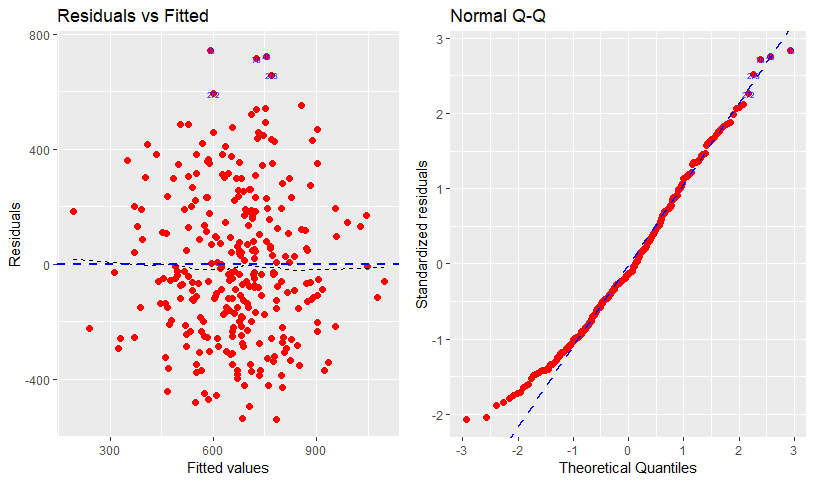
---

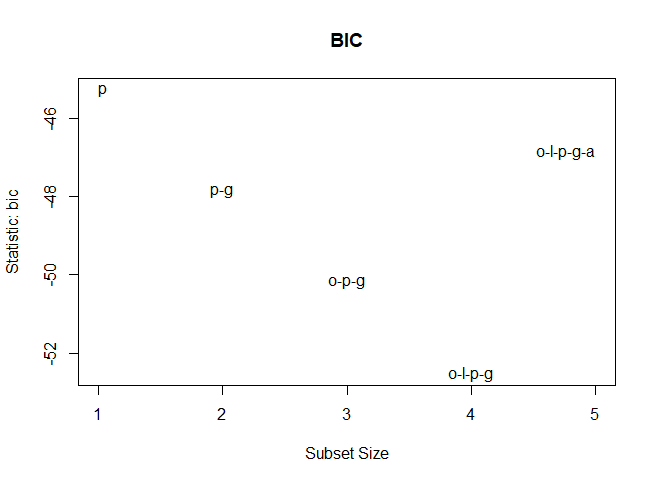
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 263.3 on 289 degrees of freedom

Multiple R-squared: 0.24, Adjusted R-squared: 0.2269

F-statistic: 18.25 on 5 and 289 DF, p-value: 9.598e-16



* fit2 Model chosen by using the BIC criterion.

for this we will select O-L-P-G as out model since it has the smallest BIC value

cwb\_fit\_2 = lm(duration ~ offer + lapse + price + gender, data = cwb)

summary(cwb\_fit\_2) # Everything is significant

> summary(cwb\_fit\_2)

Call:

lm(formula = duration ~ offer + lapse + price + gender, data = cwb)

Residuals:

Min 1Q Median 3Q Max

-539.09 -196.33 -33.47 182.37 745.29

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 838.6231 92.5116 9.065 < 2e-16 \*\*\*

offer -11.8743 3.6445 -3.258 0.001255 \*\*

lapse 1.0896 0.3853 2.828 0.005017 \*\*

price -8.3215 1.0309 -8.072 1.85e-14 \*\*\*

gender1 113.3137 31.5943 3.587 0.000393 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 262.9 on 290 degrees of freedom

Multiple R-squared: 0.24, Adjusted R-squared: 0.2295

F-statistic: 22.9 on 4 and 290 DF, p-value: < 2.2e-16

1. Use a 5-fold cross validation to compare the predictive performance of the two models. Repeat your CV for 100 times, then make a side-by-side boxplot of the delta[1] values. Generally, which model has stronger prediction ability?

d1 = numeric(100)

d2 = numeric(100)

for (i in 1:100){

d1[i] = cv.glm(cwb, cwb\_mod1, K = 5)$delta[1]

d2[i] = cv.glm(cwb, cwb\_mod2, K = 5)$delta[1]}

delta = c(d1, d2)

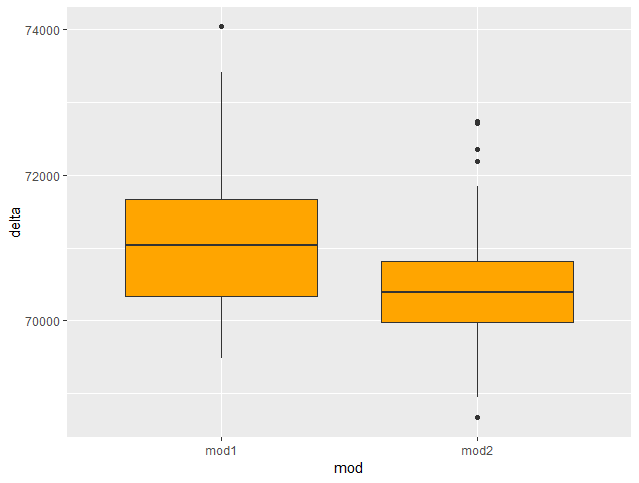
mod = rep(c("mod1", "mod2"), each = 100)

dat = tibble(delta, mod)

dat %>%

ggplot(aes(x = mod, y = delta))+

geom\_boxplot(fill = "orange")



I would interpret this as model 2 being the better pick, it has a lower delta (prediction error) values.

However, when I compare the MSE of both models:

cwb\_pred1 = predict(cwb\_mod1, cwb)

cwb\_pred2 = predict(cwb\_mod2, cwb)

cwb\_mse1 = mean((cwb$duration - cwb\_pred1)^2)

cwb\_mse2 = mean((cwb$duration - cwb\_pred2)^2)

cwb\_mse1

cwb\_mse2

> cwb\_mse1

[1] 67936.07

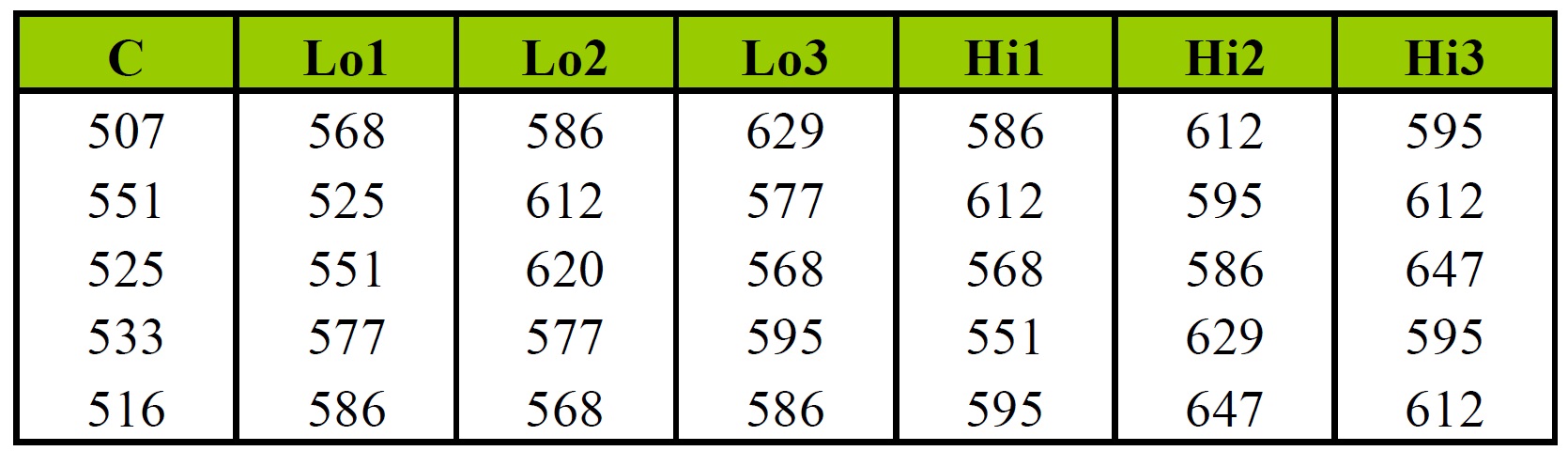
> cwb\_mse2

[1] 67936.91

…They have very similar values. Looking at the MSE wouldn’t be a good indicator of picking a better model in this case, but I thought since the mean delta in the side-by-side box plots was lower for mod2, we should also see a lower MSE.

# Problem 3 (One-way ANOVA)

The objective of the following crop management experiment is to test the effect of applied copper sulfate on grape yield by three methods: A single application for the season (1), two applications over the course of the season (2), and three applications over the course of the season (3). For each one of these application strategies, the total amount of copper sulfate was applied at two rates (Hi and Lo). A control (C) receiving no copper sulfate was also included, resulting in a total of 7 treatments (C, Lo1-3, Hi1-3) which were randomly assigned to 35 rows of vines in the an experimental vineyard. The yield data (in pounds) is presented below:



Problem 3 Data  
  
  
  
3.1 What is the response variable? and the experimental unit?

In this example the response variable is the yield of the grapes in pounds.

The experimental units are the rows of grapes.

> summary(e.mod)

Call:

lm(formula = y ~ treatment, data = cme.long)

Residuals:

Min 1Q Median 3Q Max

-32.8 -16.4 -1.4 13.9 38.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 526.400 9.845 53.468 < 2e-16 \*\*\*

treatmentHi1 56.000 13.923 4.022 0.000396 \*\*\*

treatmentHi2 87.400 13.923 6.277 8.71e-07 \*\*\*

treatmenthi3 85.800 13.923 6.162 1.18e-06 \*\*\*

treatmentLo1 31.400 13.923 2.255 0.032125 \*

treatmentLo2 66.200 13.923 4.755 5.42e-05 \*\*\*

treatmentLo3 64.600 13.923 4.640 7.42e-05 \*\*\*

---

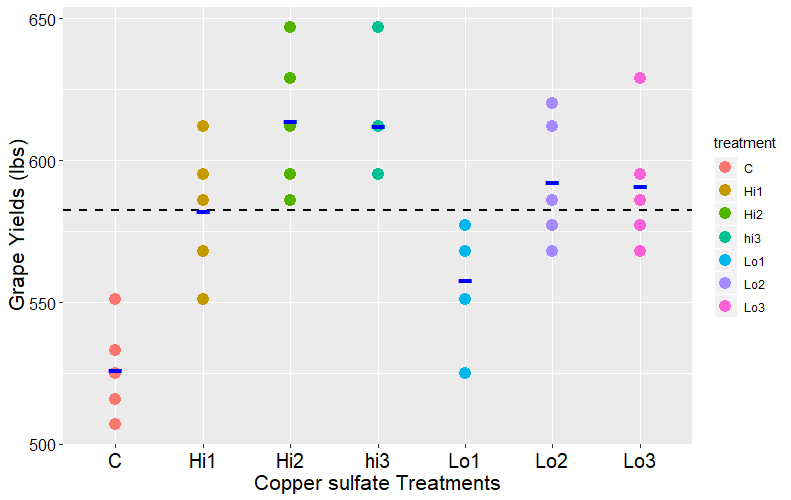
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.01 on 28 degrees of freedom

Multiple R-squared: 0.681, Adjusted R-squared: 0.6126

F-statistic: 9.961 on 6 and 28 DF, p-value: 6.7e-06

## 3.2 Visualize the data.



## 3.3 Present a table of orthogonal coefficients to answer the following questions:

> contrastmatrix

[,1] [,2] [,3] [,4] [,5] [,6]

[1,] 6 0 0 0 0 0

[2,] -1 1 2 0 2 0

[3,] -1 1 -1 1 -1 1

[4,] -1 1 -1 -1 -1 -1

[5,] -1 -1 2 0 -2 0

[6,] -1 -1 -1 1 1 -1

[7,] -1 -1 -1 -1 1 1

> grape\_contrast\_model = aov(yield ~ treatment, grape)

> grape\_contrast\_model

Call:

aov(formula = yield ~ treatment, data = grape)

Terms:

treatment Residuals

Sum of Squares 28965.54 13570.00

Deg. of Freedom 6 28

Residual standard error: 22.01461

Estimated effects are balanced

treatment 6 28966 4828 9.961 6.70e-06 \*\*\*

treatment: CS Response 1 18237 18237 37.631 1.28e-06 \*\*\*

treatment: CS Level 1 3741 3741 7.719 0.009647 \*\*

treatment: 1 vs Split 1 6955 6955 14.351 0.000739 \*\*\*

treatment: 2 vs 3 1 13 13 0.026 0.872067

treatment: 1 vs split \* CS 1 19 19 0.040 0.843403

treatment: 2 vs 3 \* CS Level 1 0 0 0.000 1.000000

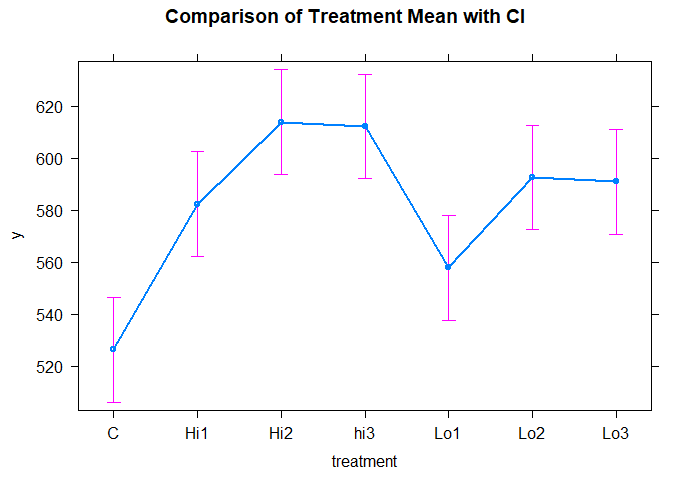
Residuals 28 13570 485

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

|  |
| --- |
| > summary(grape\_contrast\_model)  Df Sum Sq Mean Sq F value Pr(>F)  treatment 6 28966 4828 9.961 6.7e-06 \*\*\*  Residuals 28 13570 485  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| |  | | --- | |  | |

1. Is there a significant yield response to copper sulfate?
   1. It is significant with a p-value of near 0.
   2. Technically Lo1 and control have slight overlapping CI’s,
2. Is the average yield response different at the two rates (Hi and Lo)?
   1. It is significant with a p-value of near 0.



1. Are split applications superior to a single application?

The split application is also significant with a p-value of 0.01 and would work. Not as significant as the single application.

1. Considering the split applications only, is there a difference between double and triple applications?
   1. No, The double and triple applications are not significant
2. Does the application method, single or split, influence the effects of the different copper sulfate rates (Hi and Lo)?

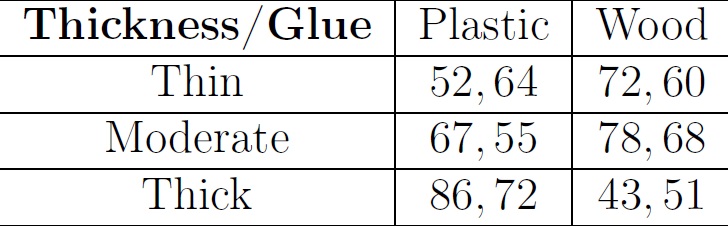
Single vs split has an effect on the different copper sulfate rates.

1. Does the number of split applications influence the effects of the different copper sulfate rates (Hi and Lo)?

No, it seems that the number of split applications increases the p-value.

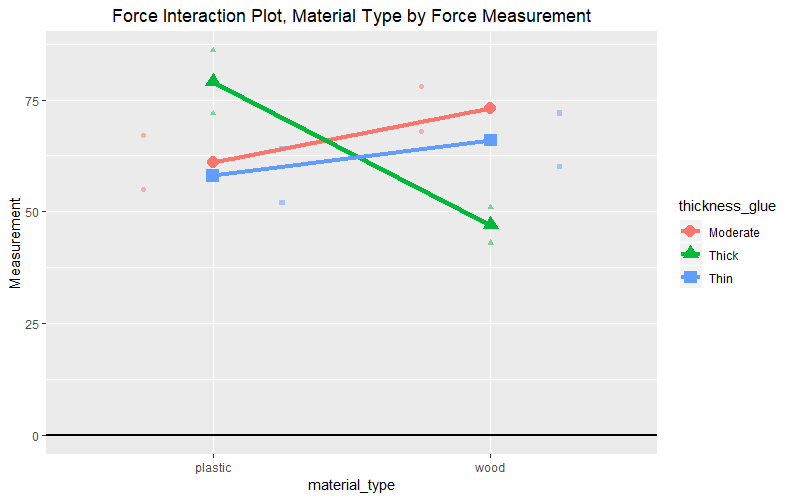
# problem 4 (Two-way ANOVA)

Consider the following small dataset, recording response variable as the amount of **force** in Newtons that it takes to separate two pieces of plastic that have been glued together, with two factors: **thickness** of the material (thin, moderate, thick), and the **type** of glue used (wood vs. plastic). There are two cases at each combination of factors.



Problem 4 Data

## 4.1 Visualize the data with one suitable plot. (Note: I want your plot to show both the distribution and interaction!)



## 

## 4.2 Write down the interaction model to fit this data.

> fit\_interaction = lm(Measurement ~ material\_type\*thickness\_glue, data = force)

> Anova(fit\_interaction, type = 3)

Anova Table (Type III tests)

Response: Measurement

Sum Sq Df F value Pr(>F)

(Intercept) 7442 1 112.7576 4.109e-05 \*\*\*

material\_type 144 1 2.1818 0.19012

thickness\_glue 516 2 3.9091 0.08187 .

material\_type:thickness\_glue 1184 2 8.9697 0.01574 \*

Residuals 396 6

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## 

## 4.3 Fit your model. Is the interaction effect significant at 5% significance level?

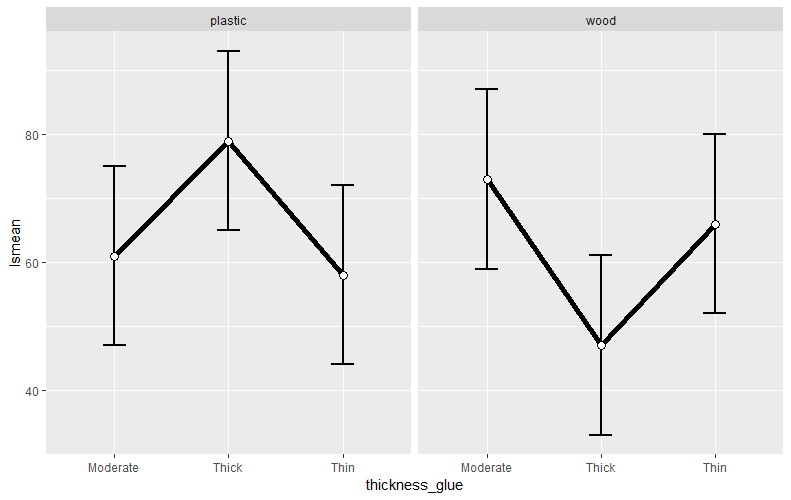
The interaction effect IS significant at the 5% significance level.

Sum Sq Df F value Pr(>F)

material\_type:thickness\_glue 1184 2 8.9697 0.01574 \*

## 4.4 Make a line plot of the Glue effects across the levels of Thickness.

# not sure why it reordered the order is Moderate > thick > thin, but it looks right compared to example in-class.



## 

## 4.5 Make a line plot of the pairwise comparison of Glue effects across the levels of Thickness.

