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

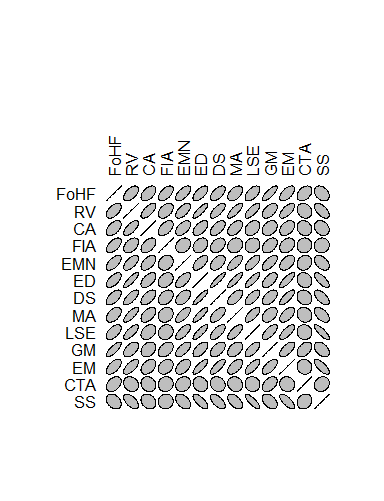
CTA: for each increase by 1 of FOFH, CTA increases by 0.159 if no other predictors change

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))

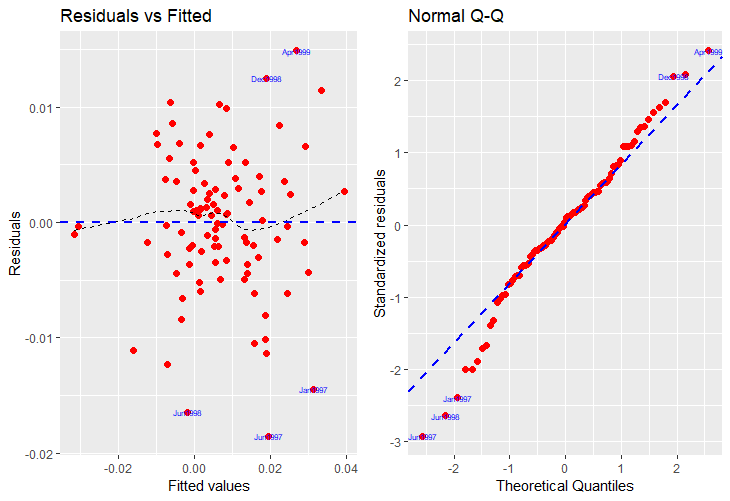


autoplot(FoHF\_fit\_1, which = 1:2, colour = 'red', size = 2,

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

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

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



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.

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.
2. Perform variable selection using the BIC criterion.

# 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.

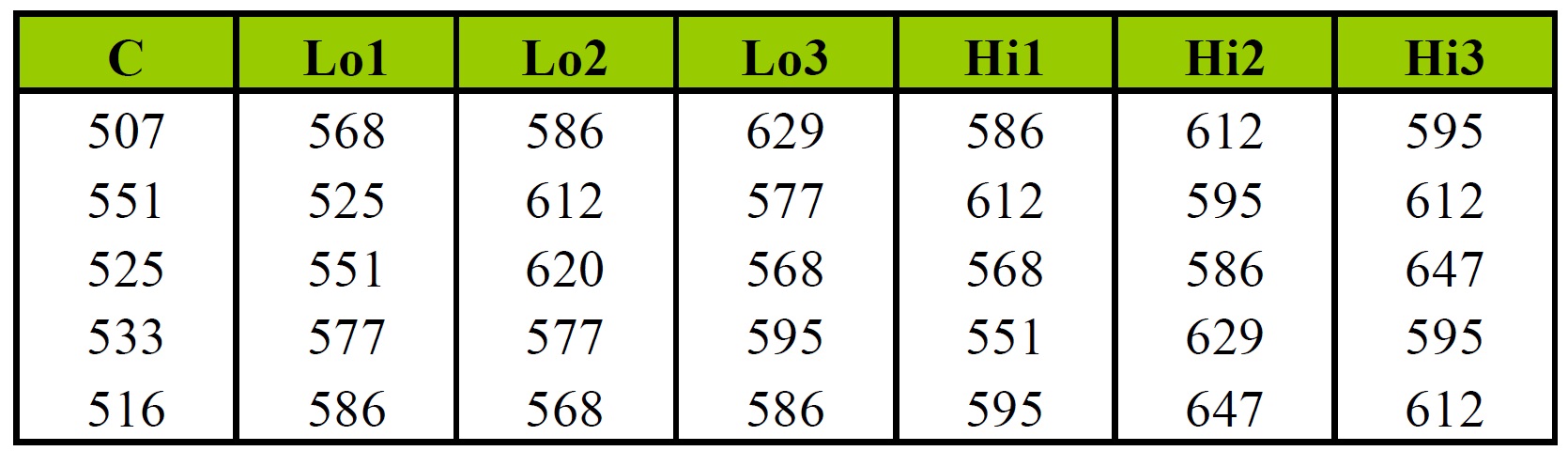
1. Change the variable gender to a factor variable and re-label its levels using “Female” and “Male”.
2. Make a scatter plot matrix of this data. Comment on any necessary patterns, variable transformations that are needed.
3. Fit the following models:

* fit1 OLS with all variables.
* fit2 Model chosen by using the BIC criterion.

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?

# Problem 3 (Onw-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?

## 3.2 Visualize the data.

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

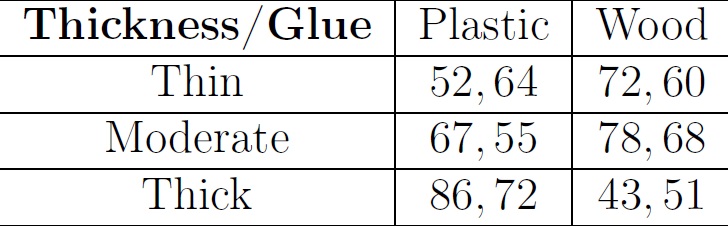
1. Is there a significant yield response to copper sulfate?
2. Is the average yield response different at the two rates (Hi and Lo)?
3. Are split applications superior to a single application?
4. Considering the split applications only, is there a difference between double and triple applications?
5. Does the application method, single or split, influence the effects of the different copper sulfate rates (Hi and Lo)?
6. Does the number of split applications influence the effects of the different copper sulfate rates (Hi and Lo)?

## 3.4 Use R to solve the contrasts you formulated.

## 3.5 Interpret all your output and write your conclusions to the six questions above based on the results.

# 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.

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

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

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