STAT 423/504 - Homework 3

Due date: Wednesday, February 26. Please submit your homework to the 423/504 mailbox at Padelford B-313 on Wednesday by 5:00 PM on the designated day. Please submit the output and plots, but not your R code unless the quesiton specifically asks for it. Total possible points: 28.

1. (7 points) 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` (on Canvas) 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

   Fit the following model:

   `FoHF` $\sim$ `RV + CA + FIA + EMN + ED + DS + MA + LSE + GM + EM + CTA + SS`

   (a) (1 point) 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 R-squared (the small p-values should indicate the indices that a FoHFs invests in)? What does a large R-squared value indicate?

   (b) (2 points) Check whether any assumptions are violated (TA and QQ plot). Also check whether there are problems with respect to multicollinearity.

   (c) (1 point) 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.
   Hint: Creating new predictors is not helpful.

   (d) (3 points) Perform variable selection using the BIC criterion. Implement the following search strategies, identify the best/final model and compare:

(i) Stepwise variable selection, starting with the full model.

(ii) Stepwise variable selection, starting with the empty model.

(iii) All Subsets variable selection.

2. (10 points) (**World cities**, ALR 8.5) The Union Bank of Switzerland publishes a report entitled "Prices and Earnings Around the Globe" on their internet web site, www.ubs.com. The data is in the file `BigMac2003` in R package alr4 and is taken from their 2003 version for 70 world cities. You can obtain a description of the data with `?BigMac2003`.

   (a) (1 point) Some predictors in this data are probably colinear or multicolinear. Based on the description of the data, which predictors are those? Print a correlation matrix to and comment on the output.

   (b) (1 point) Draw the scatterplot with `BigMac` on the vertical axis and `FoodIndex` on the horizontal axis. Provide a qualitative description of this graph.

   (c) (1 point) Use the Box–Cox method to find a transformation of `BigMac` so that the resulting scatterplot has a linear mean function.

   (d) (1 point) Two of the cities, with very large values for BigMac, are very influential for selecting a transformation. What cities are those?

   (e) (1 point) Remove the two cities you identified in the previous task and apply the Box-Cox method to the reduced data set. What has changed?

   (f) (2 points) Draw the histogram of each predictor (every variable except BigMac) in the data set. Do some of them appear right skewed? Which ones?

   (g) (1 point) Use the data where you left out the two influential points and fit the model with BigMac as the response (transformed using the Box-Cox suggested transformation) and the following predictors:

   - `log(Bread)`
   - `log(Rice)`
   - `log(Bus)`
   - `Apt`
   - `log(TeachNI)`.

   Compare the following three models in terms of their leave-one-out cross-validation score:

   1. Model with only predictors `log(Bread)`, `log(Rice)`,
   2. Model with predictors `log(Bread)`, `log(Rice)`, `Apt` and `log(Bus)`,
   3. And model with all of the predictors above in the above list.

   Which of these model acheaves the best leave-one-out cross-validation score?
   Hint: You can use the following code to obtain a leave-one-out cross-validation score.

```
# Calculate LOOCV score for a linear model
# Input: a model as fit by lm()
# Output: leave-one-out CV score
loocv.lm <- function(mdl) {
  return(mean((residuals(mdl)/(1-hatvalues(mdl)))^2))
}
```

   (h) (2 points) For the model selected in the previous task, check the model diagnostic plots (TA, QQ, Cook's distance etc.). Do you notice any model assumption violations or any unusual points?

3. Below is the partial R summary of the model $Y \sim X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_2X_3 + X_4X_5$.

```
lm(formula = Y ~ X1 + X2 * X3 + X4 * X5 + X6 + X7)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.46804    2.49497   2.993  0.00356 **
X1          -0.14175    0.45910  -0.309  0.75822
X2          -1.36912    0.60653  -2.257  0.02641 *
X3          -0.59353    0.45343  -1.309  0.19387
X4           0.03629    0.17527   0.207  0.83644
X5          -0.04930    0.04305  -1.145  0.25513
X6           0.91587    0.07335  12.487  < 2e-16 ***
X7           0.07456    0.04967   1.501  0.13681
X2:X3        1.24442    0.11245  11.066  < 2e-16 ***
X4:X5        0.20447    0.01426  14.339  < 2e-16 ***
```

(a) (1 point) Using backward elimination with p-values for $\alpha_{crit} = 0.05$ and the principle of hierarchy, which variable should be removed from the model next?

a) X_1
b) X_2
c) X_4
d) X_5

(b) (1 point) Following Preetam's suggestion, Marco decides to use the AIC criterion and the `step()` function in R for his variable selection. Below you are given a partial R output of the `step()` function where the current model is $Y \sim X_2 + X_3 + X_6$.

```
Step:  AIC=338.27
Y ~ X2 + X6 + X3

       Df Sum of Sq     RSS    AIC
+ X5    1      17.0  2701.5 339.64
- X2    1   25942.5 28660.9 571.81
+ X7    1       6.9  2711.5 340.01
<none>               2718.5 338.27
- X3    1     464.0  3182.4 352.02
+ X1    1       0.1  2718.3 340.26
- X6    1    3920.5  6639.0 425.55
```

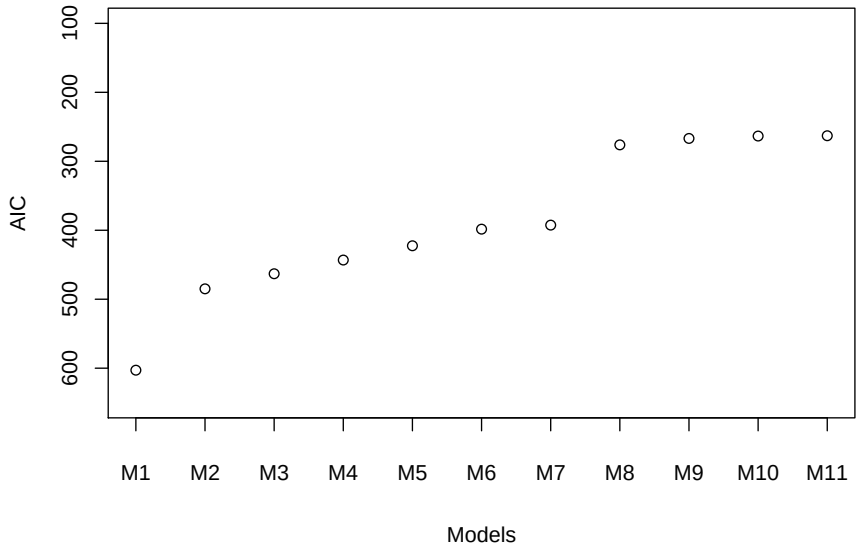
Which of the following is **true**:

a) In the next step variable X_5 will be added to the model.
b) In the next step variable X_3 will be removed from the model.
c) In the next step variable X_7 will be added to the model.
d) No variable will be added or removed in the next step.

(c) (1 point) Consider the R output in sub task b). Let model_{AIC} be the AIC of the model $Y \sim X_2 + X_6$ and let model_{BIC} be the BIC of that same model. Which of the following is **true**:

a) $\text{model}_{AIC} < \text{model}_{BIC}$
b) $\text{model}_{AIC} > \text{model}_{BIC}$
c) $\text{model}_{AIC} = \text{model}_{BIC}$
d) Cannot be answered with the provided information

(d) (1 point) The plot below shows AIC scores of all different models obtained by applying the stepwise model search in both directions to the empty model.



The models on the X-axis include the following predictors:

- M1 - none
- M2 - X_2
- M3 - X_2, X_6
- M4 - X_2, X_6, X_4
- M5 - X_2, X_6, X_4, X_3
- M6 - $X_2, X_6, X_4, X_3, X_2X_3$
- M7 - $X_2, X_6, X_4, X_3, X_5, X_2X_3$
- M8 - $X_2, X_6, X_4, X_3, X_5, X_2X_3, X_4X_5$
- M9 - $X_2, X_6, X_4, X_3, X_5, X_2X_3, X_4X_5, X_3X_6$
- M10 - $X_2, X_6, X_4, X_3, X_5, X_2X_3, X_4X_5, X_3X_6, X_2X_5$
- M11 - $X_2, X_6, X_4, X_3, X_5, X_2X_3, X_4X_5, X_3X_6, X_2X_5, X_7$

Which of the following statements is **true**?

a) The second best model in the plot contains variable X_1.

b) The best model in the plot does not contain variable X_3.

c) The worst model in the plot does not contain variable X_7.

d) The second worst model in the plot contains variable X_6.

(e) (1 point) Which of the following is **true**:

a) Given 5 predictor variables $X_1 \ldots, X_5$ all subsets regression would test 2^4 models to find the best model.

b) The BIC does not penalize the sample size.

c) The AIC penalizes the sample size.

d) The R^2 value of the model $Y \sim X_1 + X_2 + X_3$ is smaller than the R^2 value of $Y \sim X_1 + X_2 + X_3 + X_4$.

4. (5 points) (a) (1 point) Which of the following statements is **false** in the context of multiple linear regressions?

a) If the global F-test is significant, then we can conclude that $\beta_j \neq 0$ for all predictors x_j.

b) The parameters of the distribution of the test statistic that corresponds to the global F-test depend on the number of predictors and the sample size.

c) While comparing hierarchical models, we should choose the smaller model if the p-value of the corresponding F-test is larger.

d) While fitting a multiple linear regression model in `R`, the p-values provided in the summary output cannot be trusted if the errors are not normally distributed.

(b) (1 point) Which of the following statements is **true**?

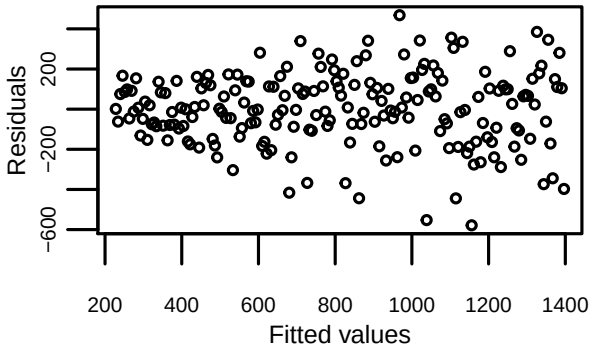
a) The p-value of a test depends on the choice of the level of significance.

b) If a 95%-confidence interval for a regression coefficient β_j contains the value 0, then the p-value for the test $H_0 : \beta_j = 0$ must be greater than 0.05.

c) If the p-value for the test $H_0 : \beta_j = 0$ is greater than 0.05, we can conclude that the corresponding predictor x_j and the response variable are uncorrelated at 95% level of significance.

d) None of the above is true.

(c) (1 point) Given the following plot, what is the most obvious model violation?



a) Correlated errors.

b) Non-constant variance of the errors.

c) Errors have a non-gaussian distribution.

d) None of the above.

(d) (1 point) Which of the following statements is **true** for a multiple linear regression model?

a) Residual analysis is not required if all predictors are statistically significant.

b) The sum of the residuals can be used to check the zero-mean assumption of the error variables.

c) The normality assumption of the errors can be checked by verifying whether the (standardized) residuals and the corresponding quantiles of a standard normal distribution have a linear relationship.

d) All data points with high leverage and small standardized residual will have a large Cook's distance.

(e) (1 point) Which of the following statements is **false** for mutiple linear regressions?

a) The statistical significance of a predictor does not imply its practical relevance.

b) Assuming that the sample size increases and predictors stay the same the individual p-values in the R summary will grow with the sample size.

c) The estimated regression coefficients can have very large standard errors in the presence of multicolinearity.

d) Multicollinearity can occur even when all pairwise correlations between the predictors are small.