Project 1 - Statistical learning

August Jonasson

2024-12-01

Task 1 - Linear regression

Loading the data.

(a)

Fitting a linear regression model with the stock log-returns as predictors and the log-return of the capital index as response.

```
model <- lm(data = returns, rOMX ~ .)
summary(model)
```

```
##
## Call:
## lm(formula = rOMX ~ ., data = returns)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0082794 -0.0006561  0.0000486  0.0007981  0.0077232
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.426e-04  3.626e-05  -3.934 8.74e-05 ***
## rABB         3.361e-02  3.839e-03   8.755  < 2e-16 ***
## rNDA.SE.ST   4.918e-02  3.625e-03  13.565  < 2e-16 ***
## HM_B.ST      4.930e-02  1.875e-03  26.291  < 2e-16 ***
## ATCO_A.ST    8.165e-02  8.631e-03   9.460  < 2e-16 ***
## ERIC_B.ST    5.023e-02  2.066e-03  24.315  < 2e-16 ***
## ESSITY_B.ST  3.230e-02  2.900e-03  11.137  < 2e-16 ***
## SAND.ST      4.660e-02  3.539e-03  13.167  < 2e-16 ***
## BOL.ST       1.792e-02  2.057e-03   8.713  < 2e-16 ***
## GETI_B.ST    1.743e-02  1.780e-03   9.792  < 2e-16 ***
## ALFA.ST      2.950e-02  2.754e-03  10.709  < 2e-16 ***
## ATCO_B.ST    3.110e-02  8.729e-03   3.563 0.000378 ***
## VOLV_B.ST    6.529e-02  3.406e-03  19.172  < 2e-16 ***
## SHB_A.ST     3.669e-02  3.783e-03   9.697  < 2e-16 ***
## ELUX_B.ST    1.188e-02  2.132e-03   5.571 3.01e-08 ***
## SEB_A.ST     4.289e-02  4.229e-03  10.142  < 2e-16 ***
## ASSA_B.ST    6.301e-02  3.278e-03  19.222  < 2e-16 ***
## AZN.ST       3.538e-02  2.666e-03  13.269  < 2e-16 ***
## SWED_A.ST    4.796e-02  3.353e-03  14.302  < 2e-16 ***
## TELIA.ST     2.910e-02  3.465e-03   8.397  < 2e-16 ***
## TEL2_B.ST    1.754e-02  2.844e-03   6.168 8.93e-10 ***
## SBB_B.ST     3.024e-03  1.005e-03   3.009 0.002667 **
```

```
## INVE_B.ST    1.074e-01  4.987e-03  21.544  < 2e-16 ***
## SINCH.ST     9.313e-03  9.680e-04   9.621  < 2e-16 ***
## SCA_B.ST     1.668e-02  2.642e-03   6.311 3.65e-10 ***
## HEXA_B.ST    5.632e-02  2.862e-03  19.681  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0014 on 1475 degrees of freedom
## Multiple R-squared:  0.9866, Adjusted R-squared:  0.9864
## F-statistic:  4341 on 25 and 1475 DF,  p-value: < 2.2e-16
```

According to the p-values of the above summary, none of the features are insignificant in their ability to predict the log-returns on the capital market index. This is not surprising at all, since we chose our predictors as the most important stocks on the market. The capital market index is modeled after these stocks. Also, the effects (coefficient estimates) of all features are very similar, i.e. no particular one feature stands out as having more or less of an impact.

By significance, we mean that under the null-hypothesis: that said coefficient has no effect on the response while keeping the others constanst, the observed value would be less than 5 % likely to occur (95 % significance level). For all of our coefficients, this probability is well below 5 % and for most of them, this probability is more or less zero.

(b)

No, the results from part (a) cannot be used to answer the question of which of the stocks have to be included in the mdoel in order to mimic the behavior of the swedish capital market index. We have not adressed potential contaminators such as multicolinearity, overfitting and joint significance between the predictors (we have only checked marginal significance). We have also not validated our model in the sense that we have no idea how it will perform on actual test data.

(c)

Now using the forward selection in order to select the model. This is done by initially only using an intercept and the response variable, and then iteratively adding whichever feature would yield the most significance until no further improvement is seen.

```
forward_model <- step(model, direction = "forward",
                      scope = formula(~ .))
```

```
## Start:  AIC=-19701.39
## rOMX ~ rABB + rNDA.SE.ST + HM_B.ST + ATCO_A.ST + ERIC_B.ST +
##     ESSITY_B.ST + SAND.ST + BOL.ST + GETI_B.ST + ALFA.ST + ATCO_B.ST +
##     VOLV_B.ST + SHB_A.ST + ELUX_B.ST + SEB_A.ST + ASSA_B.ST +
##     AZN.ST + SWED_A.ST + TELIA.ST + TEL2_B.ST + SBB_B.ST + INVE_B.ST +
##     SINCH.ST + SCA_B.ST + HEXA_B.ST
```

```
summary(forward_model)
```

```
##
## Call:
## lm(formula = rOMX ~ rABB + rNDA.SE.ST + HM_B.ST + ATCO_A.ST +
##     ERIC_B.ST + ESSITY_B.ST + SAND.ST + BOL.ST + GETI_B.ST +
##     ALFA.ST + ATCO_B.ST + VOLV_B.ST + SHB_A.ST + ELUX_B.ST +
##     SEB_A.ST + ASSA_B.ST + AZN.ST + SWED_A.ST + TELIA.ST + TEL2_B.ST +
##     SBB_B.ST + INVE_B.ST + SINCH.ST + SCA_B.ST + HEXA_B.ST, data = returns)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -0.0082794 -0.0006561  0.0000486  0.0007981  0.0077232
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.426e-04  3.626e-05  -3.934 8.74e-05 ***
## rABB         3.361e-02  3.839e-03   8.755  < 2e-16 ***
## rNDA.SE.ST   4.918e-02  3.625e-03  13.565  < 2e-16 ***
## HM_B.ST      4.930e-02  1.875e-03  26.291  < 2e-16 ***
## ATCO_A.ST    8.165e-02  8.631e-03   9.460  < 2e-16 ***
## ERIC_B.ST    5.023e-02  2.066e-03  24.315  < 2e-16 ***
## ESSITY_B.ST  3.230e-02  2.900e-03  11.137  < 2e-16 ***
## SAND.ST      4.660e-02  3.539e-03  13.167  < 2e-16 ***
## BOL.ST       1.792e-02  2.057e-03   8.713  < 2e-16 ***
## GETI_B.ST    1.743e-02  1.780e-03   9.792  < 2e-16 ***
## ALFA.ST      2.950e-02  2.754e-03  10.709  < 2e-16 ***
## ATCO_B.ST    3.110e-02  8.729e-03   3.563 0.000378 ***
## VOLV_B.ST    6.529e-02  3.406e-03  19.172  < 2e-16 ***
## SHB_A.ST     3.669e-02  3.783e-03   9.697  < 2e-16 ***
## ELUX_B.ST    1.188e-02  2.132e-03   5.571 3.01e-08 ***
## SEB_A.ST     4.289e-02  4.229e-03  10.142  < 2e-16 ***
## ASSA_B.ST    6.301e-02  3.278e-03  19.222  < 2e-16 ***
## AZN.ST       3.538e-02  2.666e-03  13.269  < 2e-16 ***
## SWED_A.ST    4.796e-02  3.353e-03  14.302  < 2e-16 ***
## TELIA.ST     2.910e-02  3.465e-03   8.397  < 2e-16 ***
## TEL2_B.ST    1.754e-02  2.844e-03   6.168 8.93e-10 ***
## SBB_B.ST     3.024e-03  1.005e-03   3.009 0.002667 **
## INVE_B.ST    1.074e-01  4.987e-03  21.544  < 2e-16 ***
## SINCH.ST     9.313e-03  9.680e-04   9.621  < 2e-16 ***
## SCA_B.ST     1.668e-02  2.642e-03   6.311 3.65e-10 ***
## HEXA_B.ST    5.632e-02  2.862e-03  19.681  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0014 on 1475 degrees of freedom
## Multiple R-squared:  0.9866, Adjusted R-squared:  0.9864
## F-statistic:  4341 on 25 and 1475 DF,  p-value: < 2.2e-16
```

(d)

```
backward_model <- step(model, direction = "backward")
```

```
## Start:  AIC=-19701.39
## rOMX ~ rABB + rNDA.SE.ST + HM_B.ST + ATCO_A.ST + ERIC_B.ST +
##     ESSITY_B.ST + SAND.ST + BOL.ST + GETI_B.ST + ALFA.ST + ATCO_B.ST +
##     VOLV_B.ST + SHB_A.ST + ELUX_B.ST + SEB_A.ST + ASSA_B.ST +
##     AZN.ST + SWED_A.ST + TELIA.ST + TEL2_B.ST + SBB_B.ST + INVE_B.ST +
##     SINCH.ST + SCA_B.ST + HEXA_B.ST
##
##               Df  Sum of Sq       RSS    AIC
## <none>                      0.0028907 -19701
## - SBB_B.ST     1 0.00001774 0.0029084 -19694
## - ATCO_B.ST    1 0.00002488 0.0029156 -19691
## - ELUX_B.ST    1 0.00006081 0.0029515 -19672
```

```
## - TEL2_B.ST    1 0.00007455 0.0029652 -19665
## - SCA_B.ST     1 0.00007806 0.0029687 -19663
## - TELIA.ST     1 0.00013817 0.0030288 -19633
## - BOL.ST       1 0.00014878 0.0030394 -19628
## - rABB         1 0.00015021 0.0030409 -19627
## - ATCO_A.ST    1 0.00017539 0.0030661 -19615
## - SINCH.ST     1 0.00018141 0.0030721 -19612
## - SHB_A.ST     1 0.00018429 0.0030750 -19611
## - GETI_B.ST    1 0.00018790 0.0030786 -19609
## - SEB_A.ST     1 0.00020158 0.0030922 -19602
## - ALFA.ST      1 0.00022476 0.0031154 -19591
## - ESSITY_B.ST  1 0.00024306 0.0031337 -19582
## - SAND.ST      1 0.00033976 0.0032304 -19537
## - AZN.ST       1 0.00034508 0.0032357 -19534
## - rNDA.SE.ST   1 0.00036060 0.0032513 -19527
## - SWED_A.ST    1 0.00040088 0.0032916 -19509
## - VOLV_B.ST    1 0.00072035 0.0036110 -19369
## - ASSA_B.ST    1 0.00072415 0.0036148 -19368
## - HEXA_B.ST    1 0.00075913 0.0036498 -19353
## - INVE_B.ST    1 0.00090959 0.0038003 -19293
## - ERIC_B.ST    1 0.00115866 0.0040493 -19198
## - HM_B.ST      1 0.00135458 0.0042453 -19127
```

```
summary(backward_model)
```

```
##
## Call:
## lm(formula = rOMX ~ rABB + rNDA.SE.ST + HM_B.ST + ATCO_A.ST +
##     ERIC_B.ST + ESSITY_B.ST + SAND.ST + BOL.ST + GETI_B.ST +
##     ALFA.ST + ATCO_B.ST + VOLV_B.ST + SHB_A.ST + ELUX_B.ST +
##     SEB_A.ST + ASSA_B.ST + AZN.ST + SWED_A.ST + TELIA.ST + TEL2_B.ST +
##     SBB_B.ST + INVE_B.ST + SINCH.ST + SCA_B.ST + HEXA_B.ST, data = returns)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0082794 -0.0006561  0.0000486  0.0007981  0.0077232
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.426e-04  3.626e-05  -3.934 8.74e-05 ***
## rABB         3.361e-02  3.839e-03   8.755  < 2e-16 ***
## rNDA.SE.ST   4.918e-02  3.625e-03  13.565  < 2e-16 ***
## HM_B.ST      4.930e-02  1.875e-03  26.291  < 2e-16 ***
## ATCO_A.ST    8.165e-02  8.631e-03   9.460  < 2e-16 ***
## ERIC_B.ST    5.023e-02  2.066e-03  24.315  < 2e-16 ***
## ESSITY_B.ST  3.230e-02  2.900e-03  11.137  < 2e-16 ***
## SAND.ST      4.660e-02  3.539e-03  13.167  < 2e-16 ***
## BOL.ST       1.792e-02  2.057e-03   8.713  < 2e-16 ***
## GETI_B.ST    1.743e-02  1.780e-03   9.792  < 2e-16 ***
## ALFA.ST      2.950e-02  2.754e-03  10.709  < 2e-16 ***
## ATCO_B.ST    3.110e-02  8.729e-03   3.563 0.000378 ***
## VOLV_B.ST    6.529e-02  3.406e-03  19.172  < 2e-16 ***
## SHB_A.ST     3.669e-02  3.783e-03   9.697  < 2e-16 ***
## ELUX_B.ST    1.188e-02  2.132e-03   5.571 3.01e-08 ***
## SEB_A.ST     4.289e-02  4.229e-03  10.142  < 2e-16 ***
```

```
## ASSA_B.ST    6.301e-02  3.278e-03  19.222  < 2e-16 ***
## AZN.ST       3.538e-02  2.666e-03  13.269  < 2e-16 ***
## SWED_A.ST    4.796e-02  3.353e-03  14.302  < 2e-16 ***
## TELIA.ST     2.910e-02  3.465e-03   8.397  < 2e-16 ***
## TEL2_B.ST    1.754e-02  2.844e-03   6.168 8.93e-10 ***
## SBB_B.ST     3.024e-03  1.005e-03   3.009 0.002667 **
## INVE_B.ST    1.074e-01  4.987e-03  21.544  < 2e-16 ***
## SINCH.ST     9.313e-03  9.680e-04   9.621  < 2e-16 ***
## SCA_B.ST     1.668e-02  2.642e-03   6.311 3.65e-10 ***
## HEXA_B.ST    5.632e-02  2.862e-03  19.681  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0014 on 1475 degrees of freedom
## Multiple R-squared:  0.9866, Adjusted R-squared:  0.9864
## F-statistic:  4341 on 25 and 1475 DF,  p-value: < 2.2e-16
```

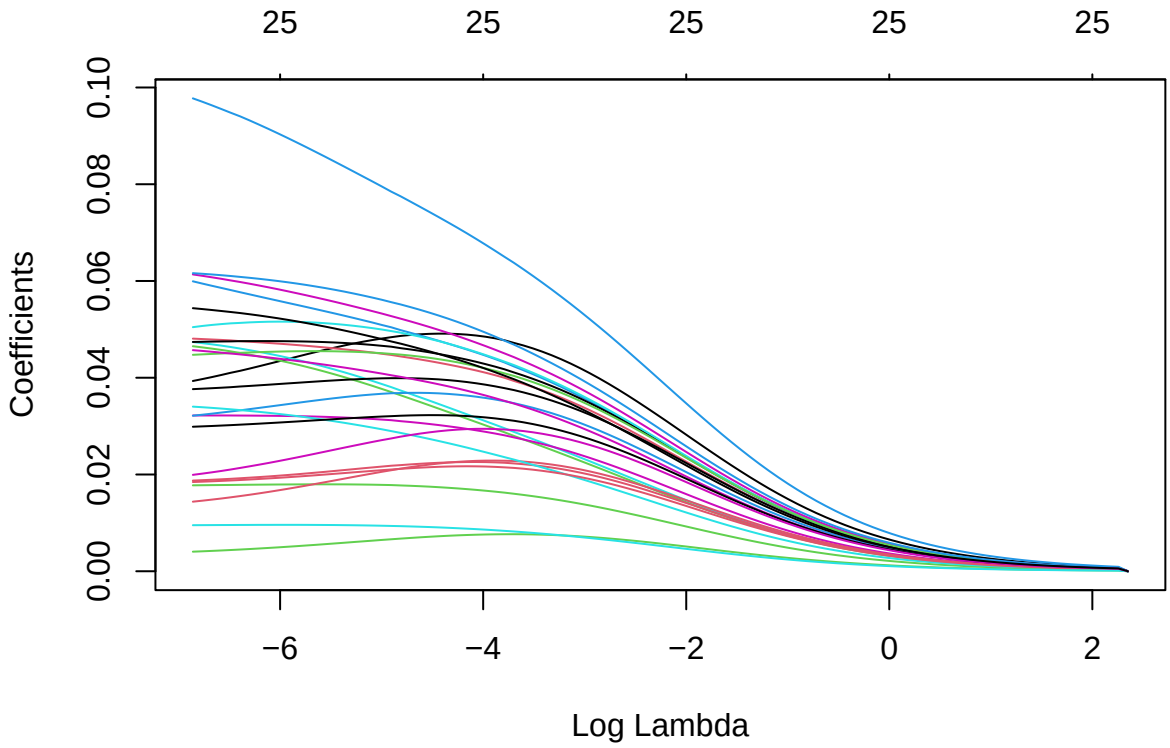
The backward selection yields the same result as the forward selection.

(e)

```
x_var <- as.matrix(returns[,2:26])
y_var <- as.matrix(returns[,1])
ridge_fit <- glmnet(x_var, y_var, alpha = 0)
summary(ridge_fit)
```

```
##           Length Class     Mode
## a0         100   -none-    numeric
## beta      2500   dgCMatrix S4
## df         100   -none-    numeric
## dim          2   -none-    numeric
## lambda     100   -none-    numeric
## dev.ratio  100   -none-    numeric
## nulldev      1   -none-    numeric
## npasses      1   -none-    numeric
## jerr         1   -none-    numeric
## offset       1   -none-    logical
## call         4   -none-    call
## nobs         1   -none-    numeric
```

```
plot(ridge_fit, xvar = "lambda", label = FALSE)
```



In order to choose the best λ we will use leave-one-out cross-validation.

```
ridge_cv <- cv.glmnet(x_var, y_var, alpha = 0)
```

(f)

(g)