Week 2 Questions

Dr M L Mackenzie

February 2018

Introduction

We will continue to use the data related to Tobin's Q which was introduced last week. These questions will be based on fitting a linear model to the data, and three alternatives: ridge regression, the LASSO and the elastic net.

Linear model

```
attach(newdat)
lmfit <- lm(tobinsQ ~ ltdratio + capexratio + rdratio + adsratio +
    pperatio + ebitdaratio + year + assets + capex + ltd + ebitda +
    ppe + sales + ads + rd + bookval + mv + indclass, data = newdat)
require(car)
vif(lmfit)
```

```
   ltdratio  capexratio     rdratio    adsratio    pperatio ebitdaratio
   1.302147    1.186232    1.338072    1.169762    1.310724    1.310487
       year      assets       capex         ltd      ebitda         ppe
   1.062863   43.364736    6.476080    9.516988   10.270968    8.376237
      sales         ads          rd     bookval          mv    indclass
   4.828356    1.967215    1.804872   12.577660    4.125966    1.070341
```

```
# center and scale covariates
xmat <- cbind(ltdratio, capexratio, rdratio, adsratio, pperatio,
    ebitdaratio, year, assets, capex, ltd, ebitda, ppe, sales,
    ads, rd, bookval, mv, indclass)
xmat <- apply(xmat, 2, scale)
summary(xmat)
```

```
    ltdratio         capexratio          rdratio           adsratio
 Min.   :-0.7680   Min.   :-2.42734   Min.   :-0.3777   Min.   :-0.46331
 1st Qu.:-0.7680   1st Qu.:-0.29652   1st Qu.:-0.3777   1st Qu.:-0.38449
 Median :-0.5049   Median :-0.18121   Median :-0.2570   Median :-0.26125
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.: 0.5645   3rd Qu.: 0.01868   3rd Qu.: 0.1042   3rd Qu.: 0.04208
 Max.   : 4.9658   Max.   :38.27773   Max.   :28.8244   Max.   :24.26125
    pperatio        ebitdaratio            year              assets
 Min.   :-1.3035   Min.   :-23.2557   Min.   :-2.10423   Min.   :-0.3418
 1st Qu.:-0.7573   1st Qu.: -0.2374   1st Qu.:-0.71465   1st Qu.:-0.3278
 Median :-0.2636   Median :  0.1347   Median : 0.05734   Median :-0.2903
 Mean   : 0.0000   Mean   :  0.0000   Mean   : 0.00000   Mean   : 0.0000
 3rd Qu.: 0.5178   3rd Qu.:  0.4792   3rd Qu.: 0.82933   3rd Qu.:-0.1241
 Max.   : 9.8381   Max.   :  7.9408   Max.   : 1.60132   Max.   :18.7325
     capex              ltd              ebitda             ppe
 Min.   :-1.6691   Min.   :-0.2644   Min.   :-2.2909   Min.   :-0.2572
 1st Qu.:-0.2543   1st Qu.:-0.2644   1st Qu.:-0.3383   1st Qu.:-0.2522
```

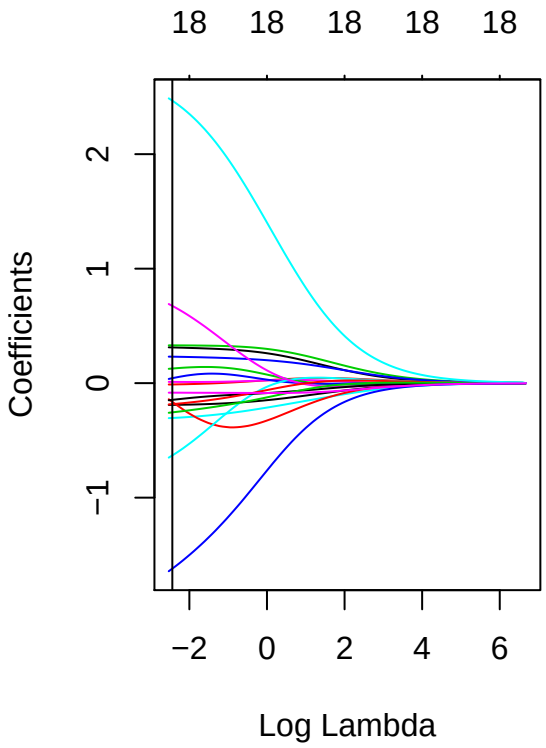
```
 Median :-0.2311   Median :-0.2616   Median :-0.3055   Median :-0.2347
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.:-0.1190   3rd Qu.:-0.1745   3rd Qu.:-0.1295   3rd Qu.:-0.1428
 Max.   :24.8298   Max.   :21.6645   Max.   :12.9302   Max.   :19.7544
     sales              ads                rd             bookval
 Min.   :-0.3271   Min.   :-0.2792   Min.   :-0.2859   Min.   :-0.36207
 1st Qu.:-0.3139   1st Qu.:-0.2755   1st Qu.:-0.2859   1st Qu.:-0.34301
 Median :-0.2754   Median :-0.2581   Median :-0.2594   Median :-0.29101
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.:-0.1037   3rd Qu.:-0.1511   3rd Qu.:-0.1315   3rd Qu.:-0.09532
 Max.   :24.0804   Max.   :19.3939   Max.   :20.6078   Max.   :22.56523
       mv             indclass
 Min.   :-0.4080   Min.   :-2.1716
 1st Qu.:-0.3940   1st Qu.:-1.0993
 Median :-0.3441   Median : 0.5093
 Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.:-0.1223   3rd Qu.: 0.6625
 Max.   :11.2310   Max.   : 1.5051
```

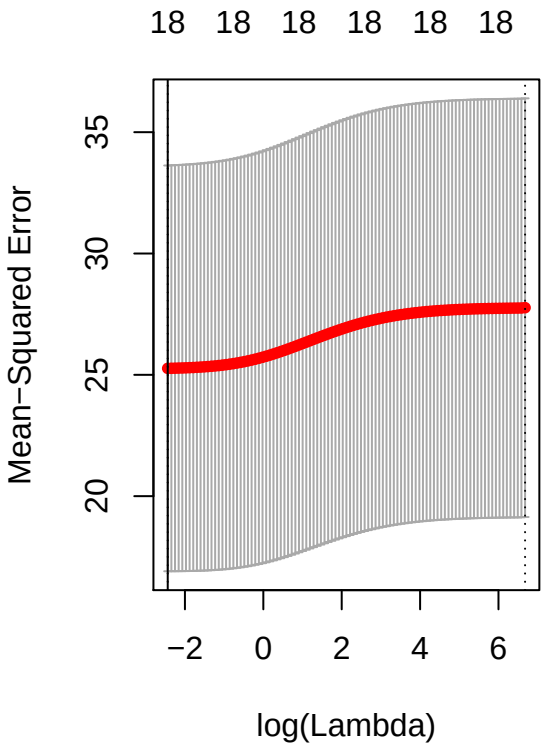
```r
# check out vifs now:
newdat3 <- data.frame(tobinsQ, xmat)
lmfit2 <- lm(tobinsQ ~ ltdratio + capexratio + rdratio + adsratio +
    pperatio + ebitdaratio + year + assets + capex + ltd + ebitda +
    ppe + sales + ads + rd + bookval + mv + indclass, data = newdat3)
require(car)
vif(lmfit2)
```

```
   ltdratio  capexratio     rdratio    adsratio    pperatio ebitdaratio
   1.302147    1.186232    1.338072    1.169762    1.310724    1.310487
       year      assets       capex         ltd      ebitda         ppe
   1.062863   43.364736    6.476080    9.516988   10.270968    8.376237
      sales         ads          rd     bookval          mv    indclass
   4.828356    1.967215    1.804872   12.577660    4.125966    1.070341
```

Ridge regression

```r
require(glmnet)

ridge <- glmnet(xmat, tobinsQ, alpha = 0)

cvridge <- cv.glmnet(cbind(ltdratio, capexratio, rdratio, adsratio,
    pperatio, ebitdaratio, year, assets, capex, ltd, ebitda,
    ppe, sales, ads, rd, bookval, mv, indclass), tobinsQ, alpha = 0,
    nfolds = 10)


par(mfrow = c(1, 2))
plot(ridge, xvar = "lambda")
abline(v = log(cvridge$lambda.min))
plot(cvridge)
abline(v = log(cvridge$lambda.min))
```

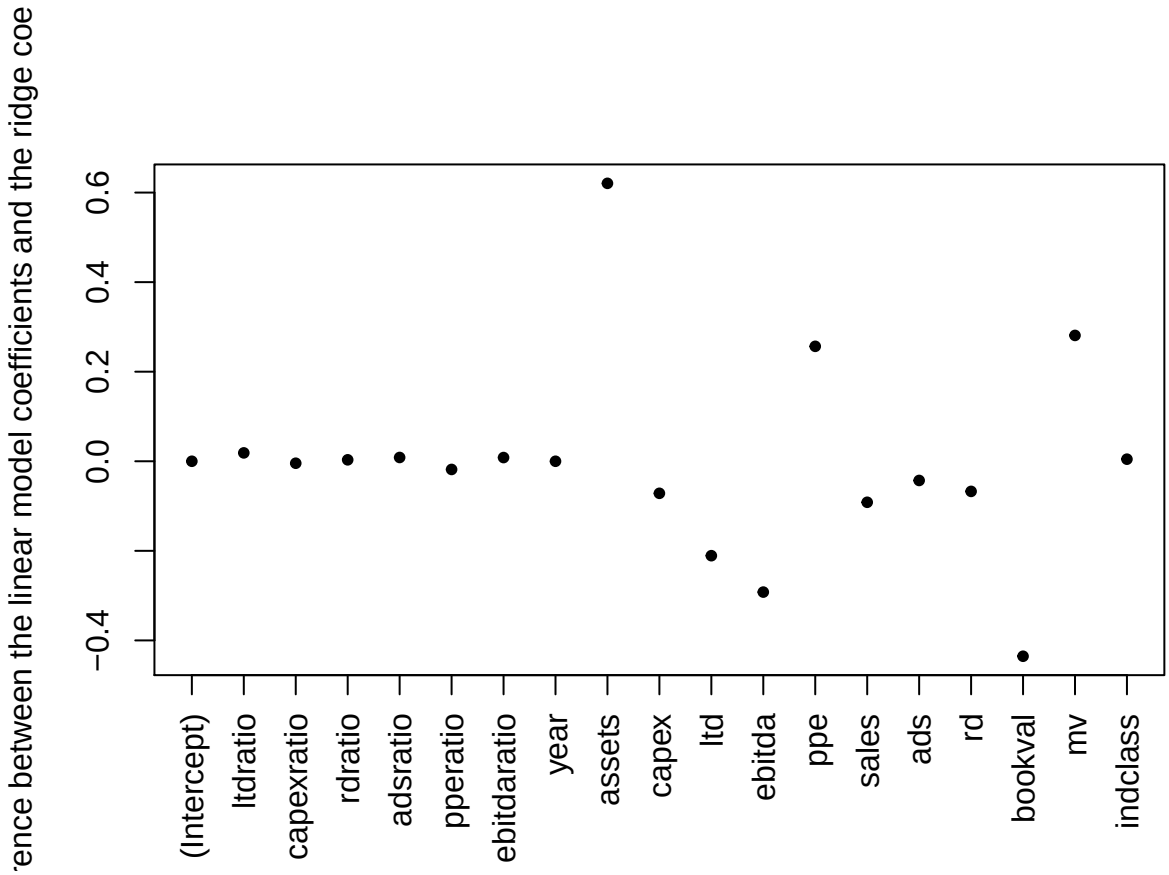




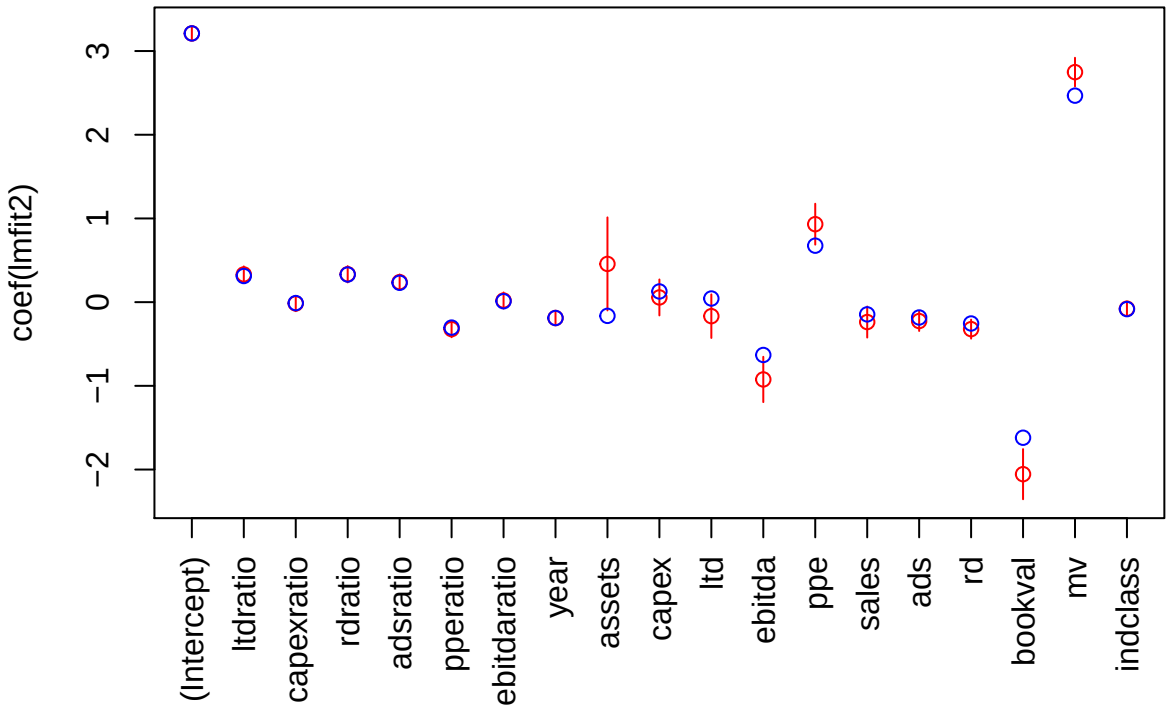

```
par(mfrow = c(1, 1))
# view the ridge regression coefficients
pickme <- which(ridge$lambda == cvridge$lambda.min)
coef(ridge)[, pickme]
```

```
(Intercept)    ltdratio  capexratio     rdratio    adsratio    pperatio
 3.21056351  0.31219563 -0.01138228  0.33005164  0.23110065 -0.30411580
ebitdaratio        year      assets       capex         ltd      ebitda
 0.01101303 -0.18965760 -0.16388299  0.12771570  0.04337725 -0.63139130
        ppe       sales         ads          rd     bookval          mv
 0.67473842 -0.14606610 -0.18328805 -0.25514899 -1.61995636  2.46704205
   indclass
-0.08314643
```

```
# plot the difference between the ridge regression
# coefficients and the lmfit2 coefficients
plot(1:19, coef(lmfit2) - coef(ridge)[, pickme], pch = 20, xaxt = "n",
    xlab = " ", ylab = "Difference between the linear model coefficients and the ridge coefficients")
axis(1, at = 1:19, labels = as.character(rownames(coef(ridge))),
    las = 2)
```




```
plot(1:19, coef(lmfit2), ylim = range(confint(lmfit2)), col = "2",
    pch = 1, xaxt = "n", xlab = " ")
segments(1:19, confint(lmfit2)[, 1], 1:19, confint(lmfit2)[,
    2], col = 2)
points(1:19, coef(ridge)[1:19, pickme], pch = 1, col = "blue")
axis(1, at = 1:19, labels = as.character(rownames(coef(ridge))),
    las = 2)
```

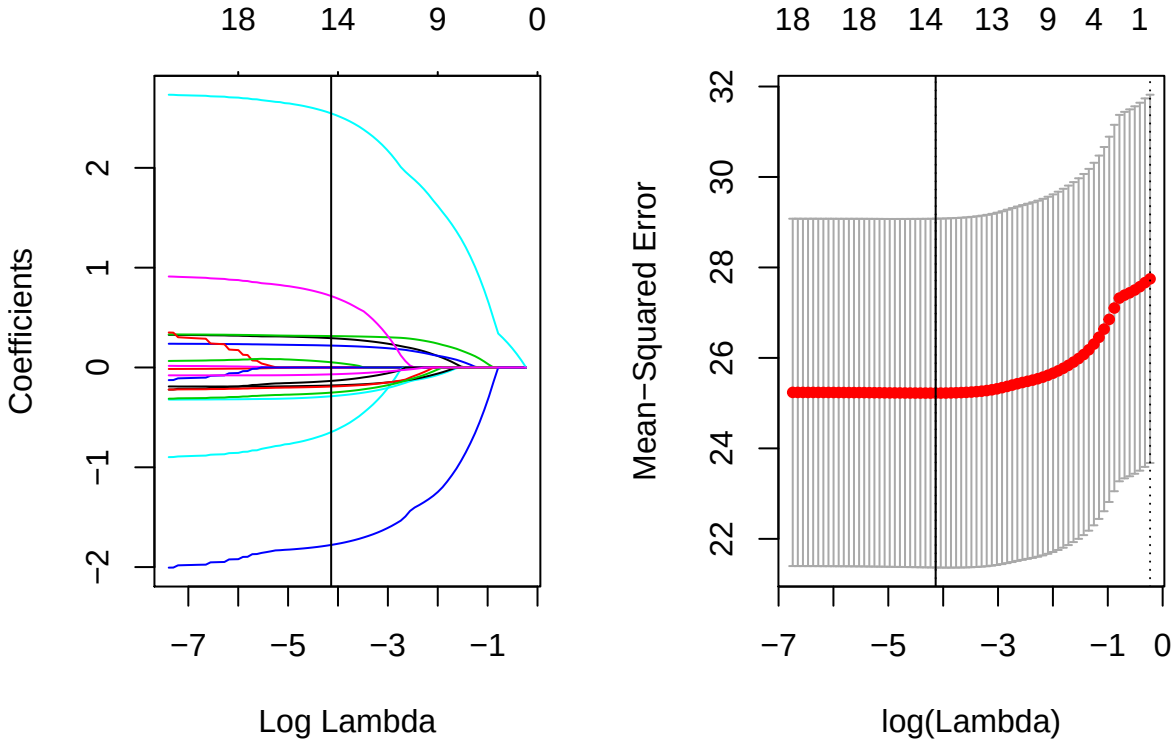



```
# confidence intervals for the lmfit2 model
confint(lmfit2)
```

```
                   2.5 %       97.5 %
(Intercept)  3.12613978  3.294987247
ltdratio     0.23455061  0.427232342
capexratio  -0.10772959  0.076176247
rdratio      0.23562690  0.430948533
adsratio     0.14825408  0.330878685
pperatio    -0.41906731 -0.225751999
ebitdaratio -0.07737085  0.115926955
year        -0.27671440 -0.102634235
assets      -0.09926351  1.012670388
capex       -0.15867647  0.271024883
ltd         -0.42794258  0.092964654
ebitda      -1.19410003 -0.652951718
ppe          0.68703761  1.175729451
sales       -0.42316616 -0.052135117
ads         -0.34453945 -0.107709521
rd          -0.43584845 -0.209001019
bookval     -2.35474826 -1.755908617
mv           2.57656361  2.919547279
indclass    -0.16581763  0.008873831
```

LASSO

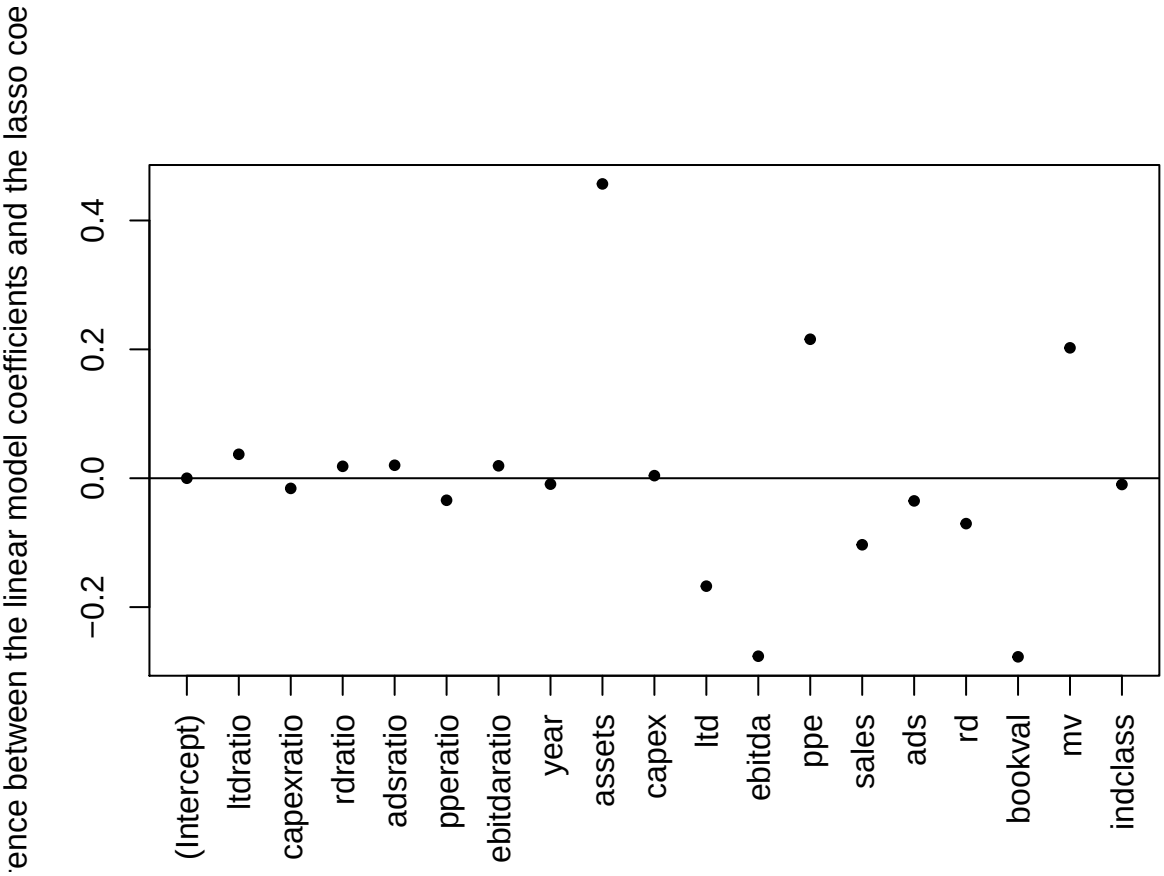
```
lasso <- glmnet(xmat, tobinsQ, alpha = 1)

cvlasso <- cv.glmnet(xmat, tobinsQ, alpha = 1, nfolds = 10)

par(mfrow = c(1, 2))
plot(lasso, xvar = "lambda")
abline(v = log(cvlasso$lambda.min))
plot(cvlasso)
abline(v = log(cvlasso$lambda.min))
```



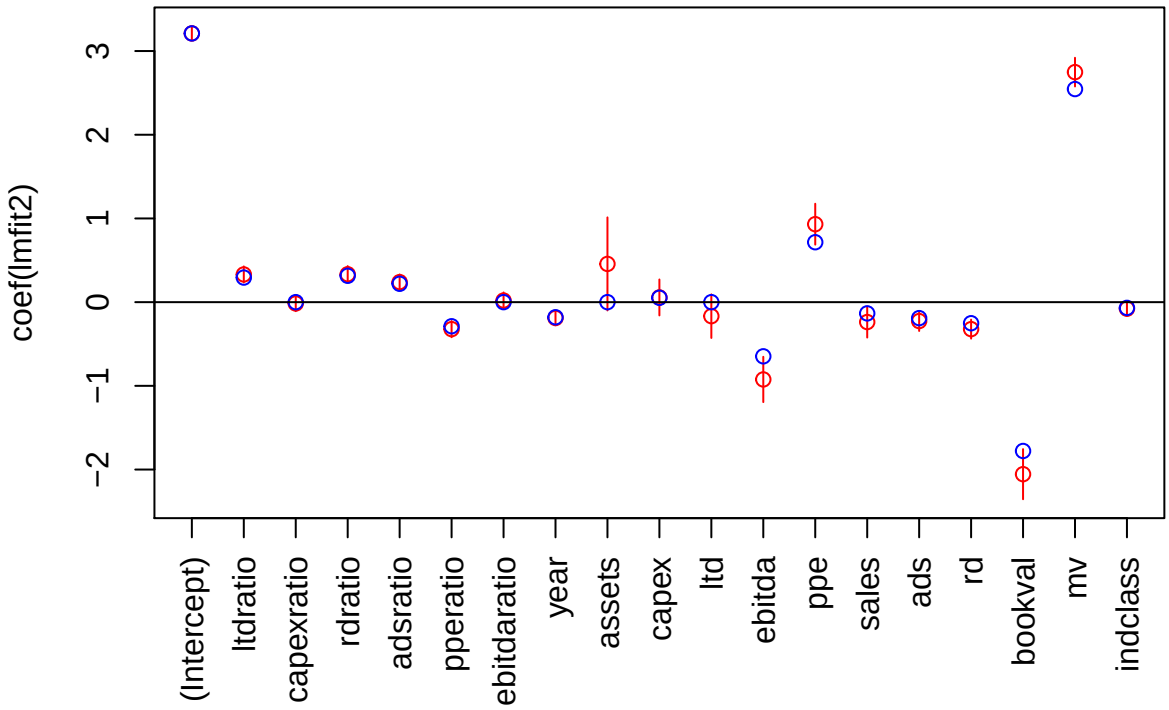

```
par(mfrow = c(1, 1))
# view the LASSO coefficients
pickme <- which(lasso$lambda == cvlasso$lambda.min)
coef(lasso)[, pickme]
```

```
(Intercept)    ltdratio  capexratio     rdratio    adsratio    pperatio
 3.21056351  0.29375121  0.00000000  0.31480775  0.21941205 -0.28822339
ebitdaratio        year      assets       capex         ltd      ebitda
 0.00000000 -0.18047129  0.00000000  0.05218285  0.00000000 -0.64734106
        ppe       sales         ads          rd     bookval          mv
 0.71572394 -0.13439922 -0.19093552 -0.25195985 -1.77819533  2.54571378
   indclass
-0.06872964
```

```
# plot the difference between the LASSO coefficients and the
# lmfit2 coefficients
plot(1:19, coef(lmfit2) - coef(lasso)[, pickme], pch = 20, xaxt = "n",
    xlab = " ", ylab = "Difference between the linear model coefficients and the lasso coefficients")
axis(1, at = 1:19, labels = as.character(rownames(coef(lasso))),
    las = 2)
abline(h = 0)
```




```
# plot the LASSO coefficients alongside the confidence
# intervals based on lmfit2.
plot(1:19, coef(lmfit2), ylim = range(confint(lmfit2)), col = "2",
    pch = 1, xaxt = "n", xlab = " ")
segments(1:19, confint(lmfit2)[, 1], 1:19, confint(lmfit2)[,
    2], col = 2)
points(1:19, coef(lasso)[1:19, pickme], pch = 1, col = "blue")
axis(1, at = 1:19, labels = as.character(rownames(coef(lasso))),
    las = 2)
abline(h = 0)
```



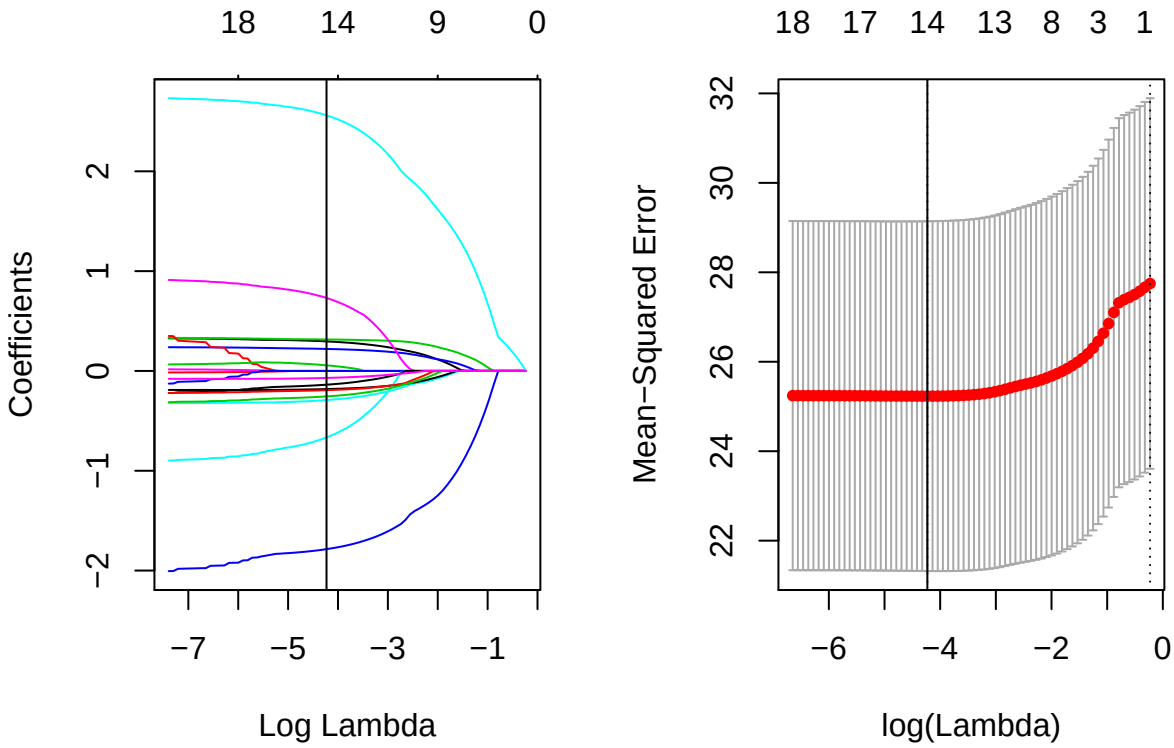

```
# confidence intervals for the lmfit2 model
confint(lmfit2)
```

```
                   2.5 %       97.5 %
(Intercept)  3.12613978  3.294987247
ltdratio     0.23455061  0.427232342
capexratio  -0.10772959  0.076176247
rdratio      0.23562690  0.430948533
adsratio     0.14825408  0.330878685
pperatio    -0.41906731 -0.225751999
ebitdaratio -0.07737085  0.115926955
year        -0.27671440 -0.102634235
assets      -0.09926351  1.012670388
capex       -0.15867647  0.271024883
ltd         -0.42794258  0.092964654
ebitda      -1.19410003 -0.652951718
ppe          0.68703761  1.175729451
sales       -0.42316616 -0.052135117
ads         -0.34453945 -0.107709521
rd          -0.43584845 -0.209001019
bookval     -2.35474826 -1.755908617
mv           2.57656361  2.919547279
indclass    -0.16581763  0.008873831
```

Elastic net

```
enet <- glmnet(xmat, tobinsQ, alpha = 0.5)

cvenet <- cv.glmnet(xmat, tobinsQ, nfolds = 10, alpha = 0.5)
```

```
par(mfrow = c(1, 2))
plot(enet, xvar = "lambda")
abline(v = log(cvenet$lambda.min))
plot(cvenet)
abline(v = log(cvenet$lambda.min))
```

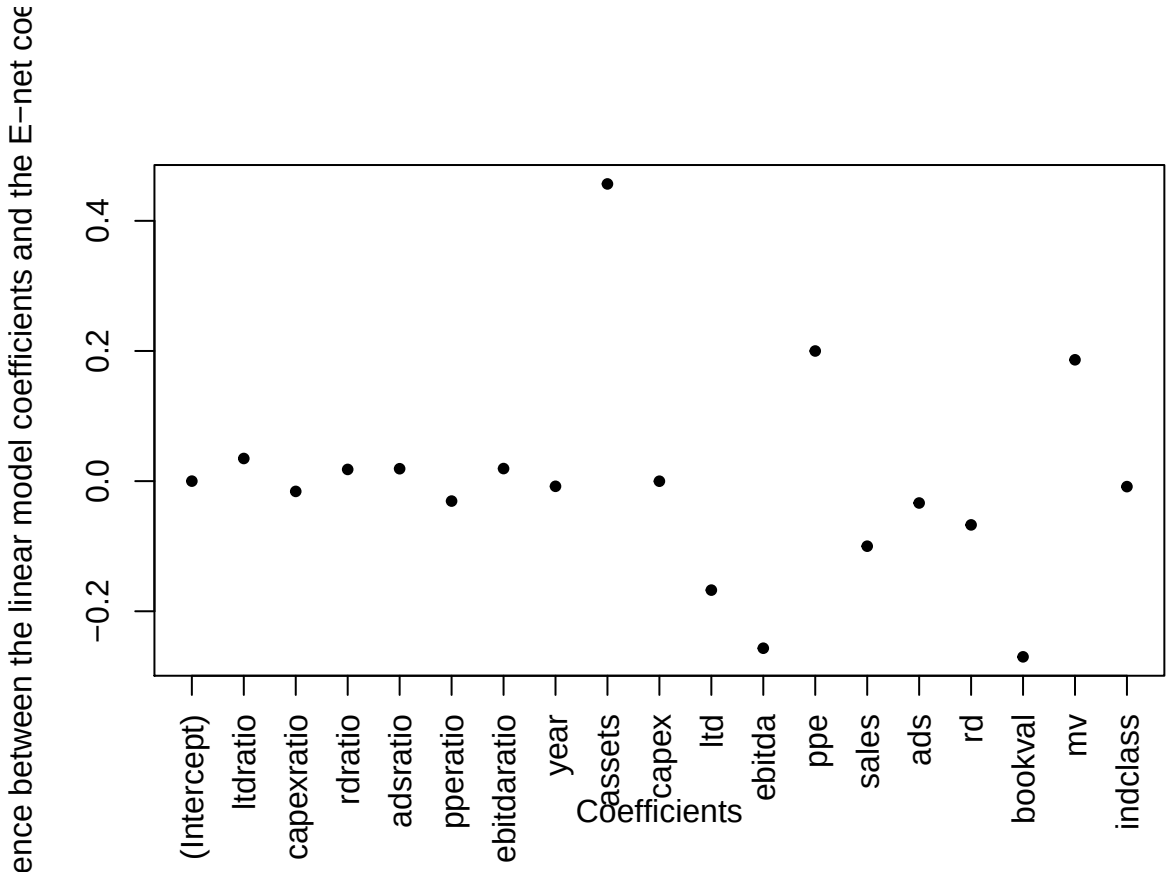



```
# check out the coefficients
pickme <- which(enet$lambda == cvenet$lambda.min)
coef(enet)[, pickme]
```

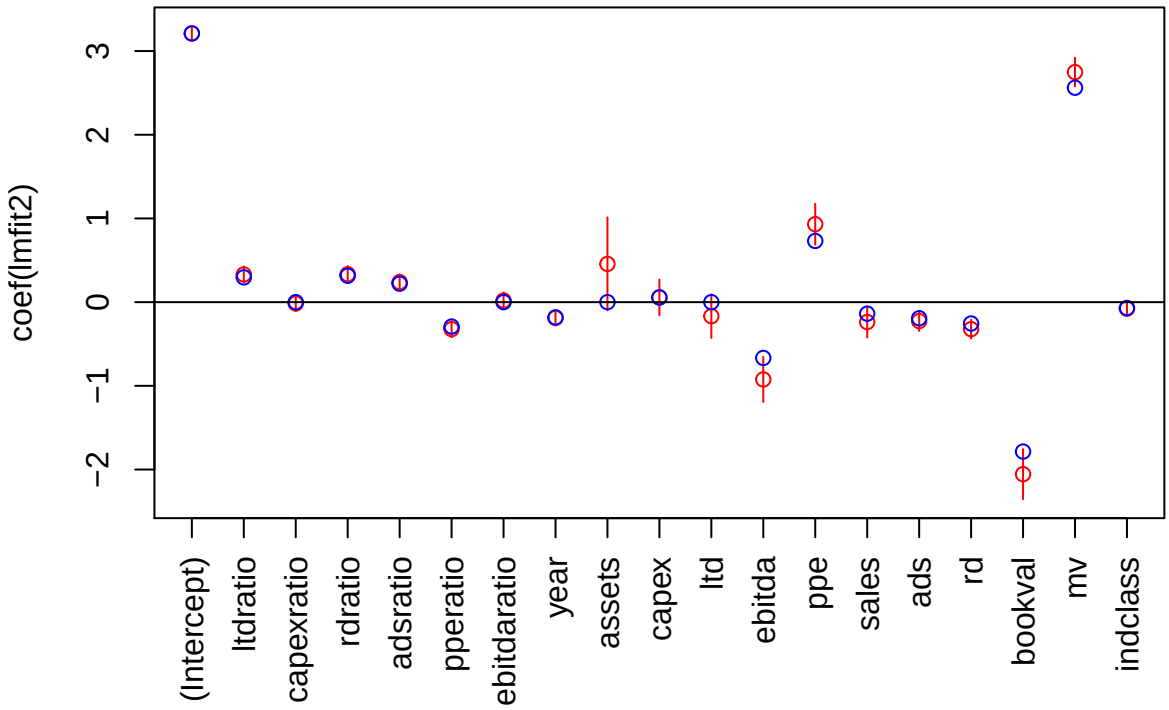
```
(Intercept)    ltdratio  capexratio     rdratio    adsratio    pperatio
 3.21056351  0.29611913  0.00000000  0.31531816  0.22050686 -0.29175412
ebitdaratio        year      assets       capex         ltd      ebitda
 0.00000000 -0.18170537  0.00000000  0.05631306  0.00000000 -0.66673818
        ppe       sales         ads          rd     bookval          mv
 0.73142568 -0.13765570 -0.19251001 -0.25515316 -1.78531088  2.56162468
   indclass
-0.07000019
```

```
par(mfrow = c(1, 1))
plot(1:19, coef(lmfit2) - coef(enet)[, pickme], pch = 20, xaxt = "n",
```

```
    xlab = "Coefficients", ylab = "Difference between the linear model coefficients and the E-net coeffi
axis(1, at = 1:19, labels = as.character(rownames(coef(enet))),
    las = 2)
```



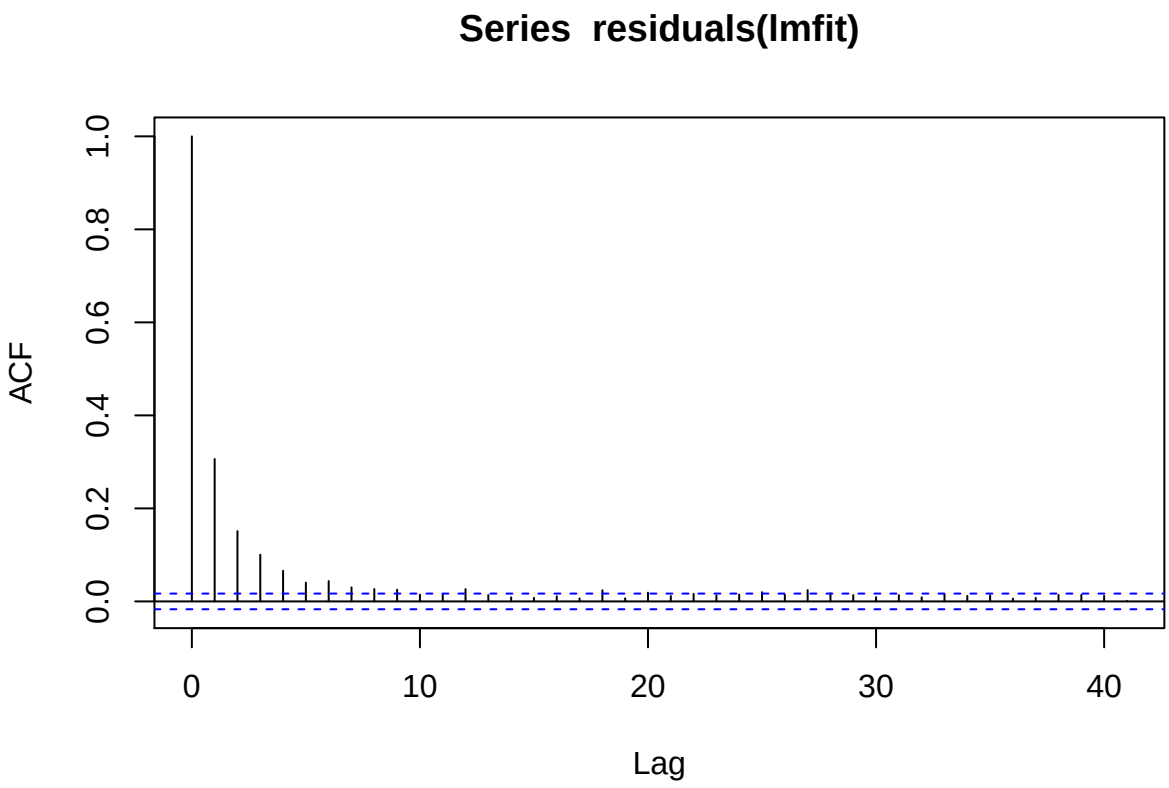

```
# plot the LASSO coefficients alongside the confidence
# intervals based on lmfit2.
plot(1:19, coef(lmfit2), ylim = range(confint(lmfit2)), col = "2",
    pch = 1, xaxt = "n", xlab = " ")
segments(1:19, confint(lmfit2)[, 1], 1:19, confint(lmfit2)[,
    2], col = 2)
points(1:19, coef(enet)[1:19, pickme], pch = 1, col = "blue")
axis(1, at = 1:19, labels = as.character(rownames(coef(enet))),
    las = 2)
abline(h = 0)
```




```
library(lawstat)
runs.test(residuals(lmfit))
```

```
    Runs Test - Two sided

data:  residuals(lmfit)
Standardized Runs Statistic = -61.543, p-value < 2.2e-16
```

```
acf(residuals(lmfit))
```



Questions

1. Which of the following is FALSE?
   - The VIF values for `assets`, `capex`, `ltd`, `ebitda`, `ppe`, `sales`, `bookval` and `mv` are at a level to cause concern.
   - Based on the `vif(lmfit)` output we can see that a linear model which has `ebitda` as the response and remaining covariates as model covariates returns an unadjusted R^2 value of 0.9026.
   - When variables are centered and scaled the coefficients need to be interpreted considering the centering and scaling which has occurred.
   - We can see from the summary of the newly created `xmat` that the mean for each covariate is now 0, however we cannot tell from this output alone if the standard deviation of each covariate is equal to 1.
   - Centering and scaling the covariates reduced the VIFs for all covariates sufficiently so that any concerns we initially had have been removed.

2. Which of the following is FALSE?
   - The absolute value of the difference between the ridge regression based coefficients (estimated using a λ chosen using CV) and the linear model coefficients (for the centered and scaled covariates) was largest for `assets` while the second largest was associated with `bookval`. This is unsurprising since these covariates have the two highest VIFs, even after centering and scaling.
   - The ridge regression parameter estimates associated with 5 of the model covariates are not located inside the 95% confidence intervals based on the `lmfit2` model.
   - All of the λ values trialled for the ridge regression were within one standard error of the average CV score obtained under the 'best' model chosen using CV.

- All 18 covariates were retained in the ridge regression model for all the candidate λ values tried in this case. Had larger λ values been trialled, the number of covariates included in the model (i.e. with non-zero valued coefficients) may have been reduced.
- The dotted line on the plot for the ridge regression showing the `log(Lambda)` versus the `Mean Squared Error` indicates `lambda.1se`. This value represents the CV value for the most penalised model which has a CV score which lies within one standard error of the smallest CV score.

3. Which of the following is FALSE?

   - Of those covariates which had zero valued coefficients returned under the LASSO (with the chosen value of λ), only `assets` and `ltd` demonstrated coefficients under the ridge regresion which were notably different (e.g. a difference > 0.1) compared with the linear model coefficients.
   - The best LASSO model (based on λ chosen using CV) returns a model with 3 less covariates (i.e. these are zero valued) than the model with all of the candidate covariates.
   - All of the λ values trialled for the LASSO were within one standard error of the average CV score obtained under the 'best' model chosen using CV. This illustrates the variability in the CV scores returned when a given value of λ is trialled.
   - While the LASSO returns zero-valued estimates for some covariates, we cannot be sure this process has 'selected' the right covariates since the LASSO does this relatively arbitrarily from a group of collinear covariates.
   - In all models fitted to date (including the `lmfit2` model), the treatment of `indclass` has been inappropriate since it has just one coefficient associated with it and this is a factor variable.

4. Which of the following is FALSE?

   - The best elastic net-based model (based on λ chosen using CV) estimated zero-valued coefficients for the same covariates as the LASSO. This is reassuring since the elastic net has a 'grouping' feature which the LASSO does not.
   - All of the elastic net parameter estimates were located inside the 95% confidence intervals based on the `lmfit2` model, suggesting the results for this procedure were more similar to the `lmfit2` results compared with the ridge regression.
   - The estimates returned for the elastic net are more similar to the corresponding estimates based on the LASSO compared with estimates based on the ridge regression. The latter produced estimates which were most disimilar to `lmfit2` but did not zero value any of the coefficients.
   - In order to obtain reliable confidence intervals about the ridge regression, LASSO or elastic net coefficients one can use the t-distribution as the basis to build confidence intervals and calculate p-values associated with each parameter. 没有distribution可以得到

5. Which of the following is FALSE?

   - The ACF plot for the residuals resulting from the `lmfit` model demonstrate positive correlation which decays with the gap/distance between observations.
   - The runs test statistic confirms that this correlation is positive since more runs were observed, compared with the number expected under the null hypothesis of independence.
   - The runs test uses a standard Normal distribution as a basis for calculating the p-value.
   - As it stands, the `lmfit` model may have standard errors and p-values which are too small and should not be trusted.