Week 4 Questions

Dr M L Mackenzie

February 2018

Introduction

We are going to use the same data as for the last three weeks of the 'Weekly Questions' to look at penalised regression splines and smoothing splines.

Penalised regression splines

```
library(mgcv)

penreg <- gam(tobinsQ ~ s(ltdratio) + s(capexratio) + s(rdratio) +
    s(adsratio) + s(pperatio) + s(ebitdaratio) + s(year) + s(assets) +
    s(capex) + s(ltd) + s(ebitda) + s(ppe) + s(sales) + s(ads) +
    s(rd) + s(bookval) + s(mv) + as.factor(indclass), data = newdat)
```

```
summary(penreg)
```

```
Family: gaussian
Link function: identity

Formula:
tobinsQ ~ s(ltdratio) + s(capexratio) + s(rdratio) + s(adsratio) +
    s(pperatio) + s(ebitdaratio) + s(year) + s(assets) + s(capex) +
    s(ltd) + s(ebitda) + s(ppe) + s(sales) + s(ads) + s(rd) +
    s(bookval) + s(mv) + as.factor(indclass)

Parametric coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.97545    0.90685   3.281  0.00104 **
as.factor(indclass)2  0.38181    0.94847   0.403  0.68728
as.factor(indclass)3  0.33885    1.53323   0.221  0.82509
as.factor(indclass)4  1.44997    1.14101   1.271  0.20383
as.factor(indclass)5  1.92312    1.56106   1.232  0.21800
as.factor(indclass)6 -0.18385    0.94615  -0.194  0.84593
as.factor(indclass)7 -1.14001    0.98698  -1.155  0.24809
as.factor(indclass)8  0.22836    1.14608   0.199  0.84207
as.factor(indclass)9 -0.15114    0.94082  -0.161  0.87237
as.factor(indclass)10 -0.04264    0.97144  -0.044  0.96499
as.factor(indclass)11 -0.21968    0.99296  -0.221  0.82491
as.factor(indclass)12  0.27309    0.92023   0.297  0.76665
as.factor(indclass)13  1.10754    0.92083   1.203  0.22909
as.factor(indclass)14  0.25694    0.95491   0.269  0.78788
as.factor(indclass)15 -0.50483    1.03282  -0.489  0.62500
as.factor(indclass)16 -1.27077    1.71043  -0.743  0.45752
as.factor(indclass)17  0.10936    0.95600   0.114  0.90893
```

```
as.factor(indclass)18 -1.03708    1.37458  -0.754  0.45058
as.factor(indclass)19  0.62789    1.12794   0.557  0.57776
as.factor(indclass)20 -0.37853    1.79171  -0.211  0.83268
as.factor(indclass)21  0.33756    0.92662   0.364  0.71565
as.factor(indclass)22 -0.30362    0.93226  -0.326  0.74467
as.factor(indclass)23  0.12028    0.95119   0.126  0.89938
as.factor(indclass)24 -0.42852    1.48723  -0.288  0.77325
as.factor(indclass)25  0.43832    1.54983   0.283  0.77732
as.factor(indclass)26  0.33545    1.09161   0.307  0.75862
as.factor(indclass)28  0.61837    1.29419   0.478  0.63280
as.factor(indclass)30 -1.19164    2.01094  -0.593  0.55347
as.factor(indclass)32  0.15467    0.97361   0.159  0.87378
as.factor(indclass)33  0.53984    1.08385   0.498  0.61844
as.factor(indclass)34  0.78071    0.94932   0.822  0.41087
as.factor(indclass)35  0.26349    0.92364   0.285  0.77544
as.factor(indclass)36  0.82897    0.91297   0.908  0.36390
as.factor(indclass)37  0.03361    0.91613   0.037  0.97073
as.factor(indclass)38  0.08538    0.92634   0.092  0.92657
as.factor(indclass)39 -0.73570    1.00669  -0.731  0.46491
as.factor(indclass)41  1.11948    1.60161   0.699  0.48458
as.factor(indclass)42 -0.08984    0.93944  -0.096  0.92381
as.factor(indclass)43 -0.05152    0.91619  -0.056  0.95516
as.factor(indclass)44 -0.24435    0.93011  -0.263  0.79278
as.factor(indclass)49 -0.95334    1.44303  -0.661  0.50885
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df       F  p-value
s(ltdratio)   7.380  8.356  23.949  < 2e-16 ***
s(capexratio) 1.000  1.000   6.961 0.008341 **
s(rdratio)    6.395  7.553   1.963 0.048434 *
s(adsratio)   7.032  7.987   4.931 4.34e-06 ***
s(pperatio)   5.940  7.116   1.667 0.111346
s(ebitdaratio) 7.897  8.693  37.728  < 2e-16 ***
s(year)       1.000  1.000   6.128 0.013314 *
s(assets)     8.257  8.818   8.569 1.28e-12 ***
s(capex)      1.000  1.000   2.666 0.102559
s(ltd)        8.265  8.838   7.645 3.77e-11 ***
s(ebitda)     5.619  6.885   9.908 3.78e-12 ***
s(ppe)        2.479  3.171   9.785 1.18e-06 ***
s(sales)      6.514  7.620   2.630 0.015886 *
s(ads)        4.778  5.854   3.465 0.002186 **
s(rd)         1.594  1.988   9.025 0.000128 ***
s(bookval)    9.000  9.000 142.839  < 2e-16 ***
s(mv)         8.946  8.999 346.704  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.323   Deviance explained = 32.9%
GCV = 18.993  Scale est. = 18.805     n = 13525
```

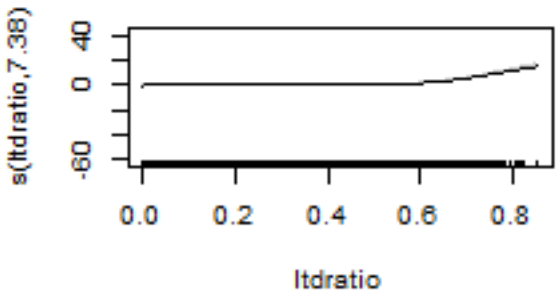
```
penreg$sp
```

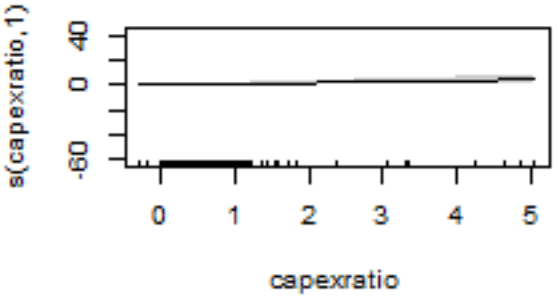
dispersion parameter

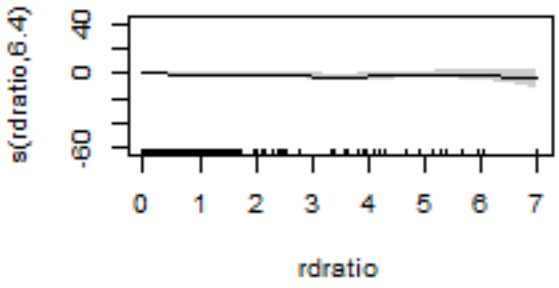
```
   s(ltdratio)  s(capexratio)     s(rdratio)    s(adsratio)    s(pperatio)
```

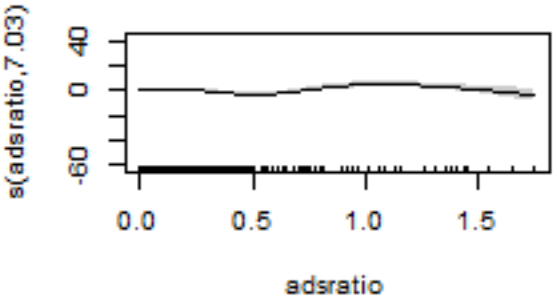
```
  2.512134e-01   4.379502e+09   8.798330e-01   1.709061e-02   3.380064e-01
s(ebitdaratio)        s(year)      s(assets)       s(capex)         s(ltd)
  1.954107e-02   3.471785e+11   5.149663e-03   1.474921e+09   1.262447e-02
     s(ebitda)         s(ppe)      s(sales)         s(ads)         s(rd)
  9.360917e-02   9.170058e-01   6.989224e-02   2.353797e-01   2.635078e+01
     s(bookval)          s(mv)
  4.221977e-12   1.683048e-03
```

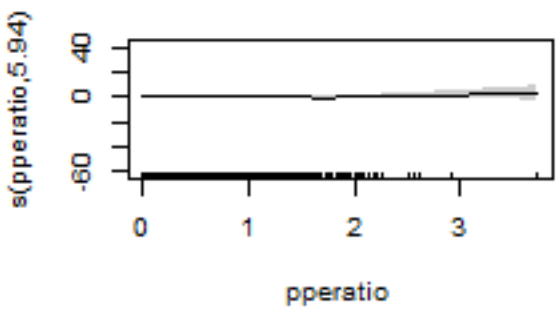
```
par(mfrow = c(2, 2))
plot(penreg, shade = T)
```

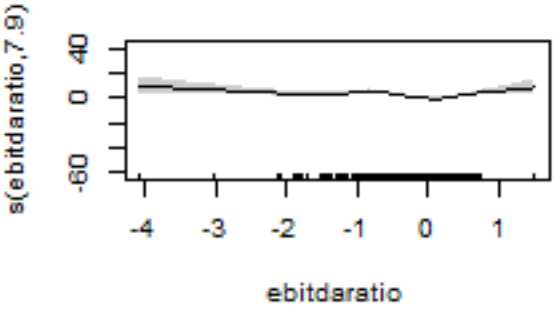


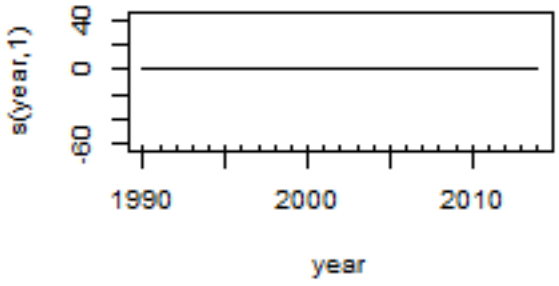


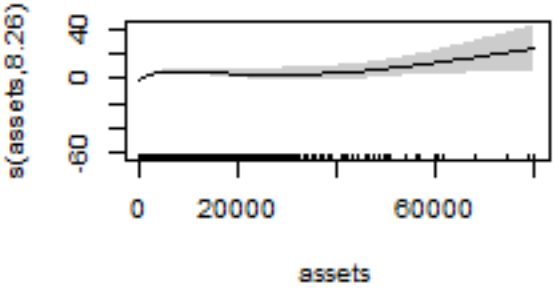


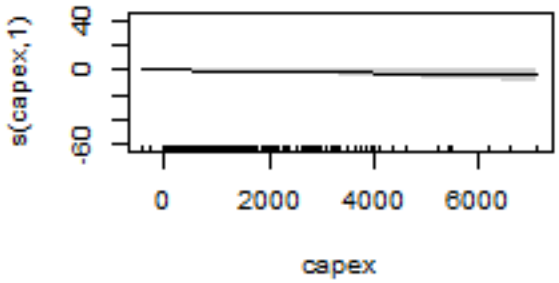


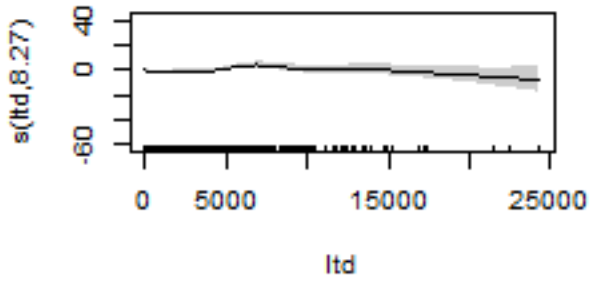

s(pperatio,5.94)
40
0
-60
0
1
2
3
pperatio

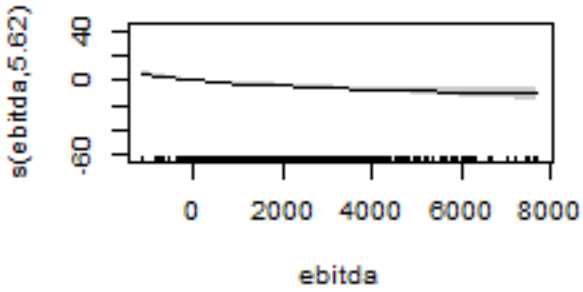

s(ebitdaratio,7.9)
40
0
-60
-4
-3
-2
-1
0
1
ebitdaratio

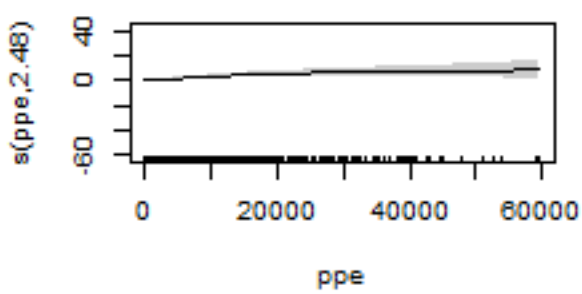

s(year,1)
40
0
-60
1990
2000
2010
year

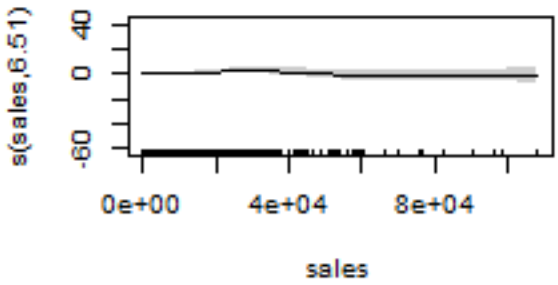

s(assets,8.26)
40
0
-60
0
20000
60000
assets

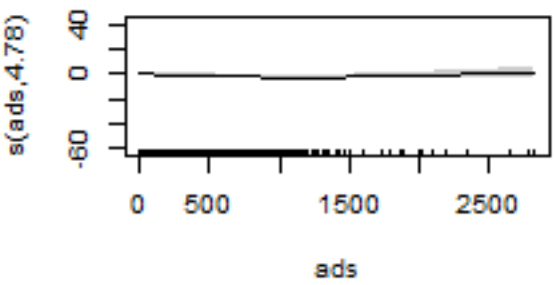

s(capex,1)
40
0
-60
0
2000
4000
6000
capex

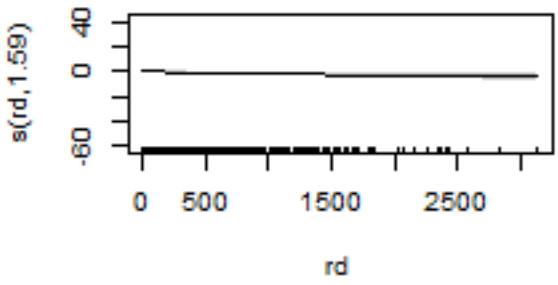

s(ltd,8.27)
40
0
-60
0
5000
15000
25000
ltd

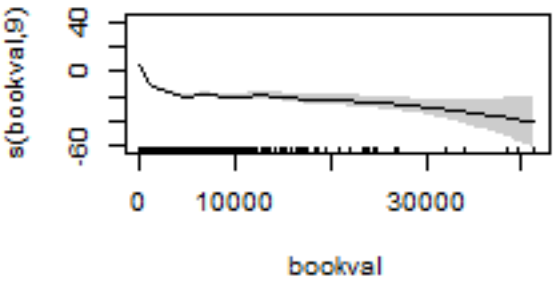

s(ebitda,5.62)
40
0
-60
0
2000
4000
6000
8000
ebitda

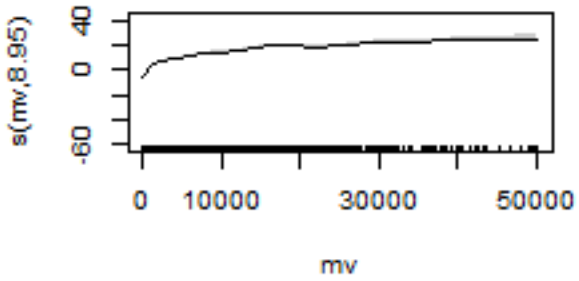

s(ppe,2.48)
40
0
-60
0
20000
40000
60000
ppe


s(sales,6.51)
40
0
-60
0e+00
4e+04
8e+04
sales


s(ads,4.78)
40
0
-60
0
500
1500
2500
ads

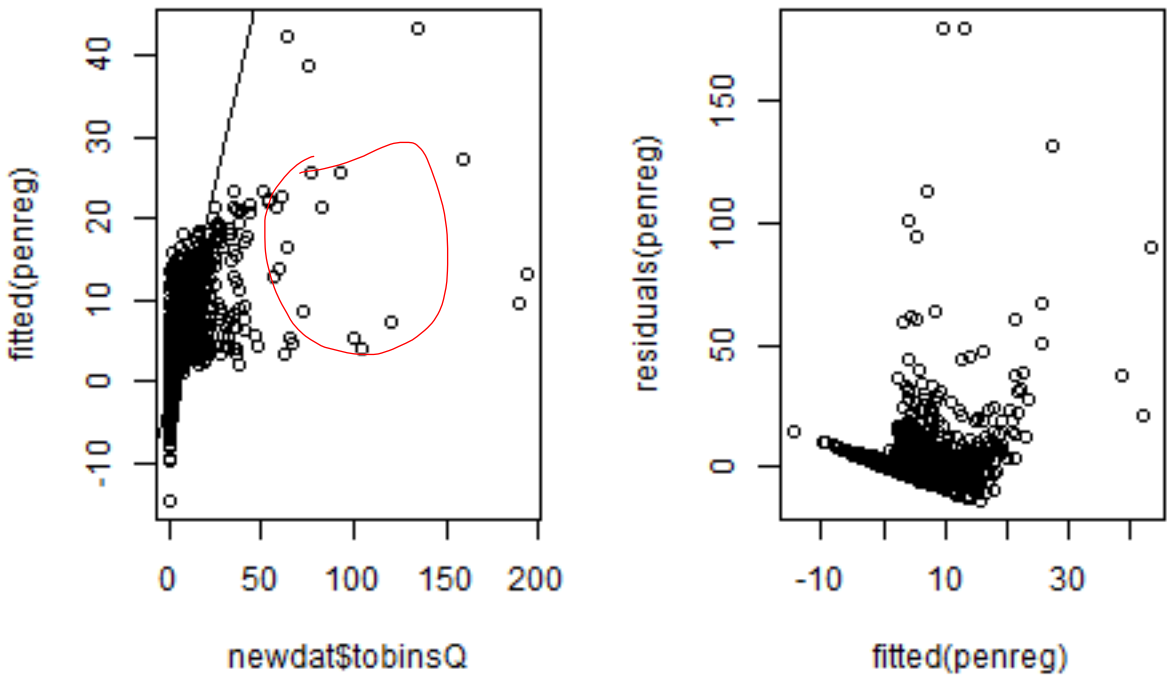

s(rd,1.59)
40
0
-60
0
500
1500
2500
rd


s(bookval,9)
40
0
-60
0
10000
30000
bookval




```
par(mfrow = c(1, 2))
plot(newdat$tobinsQ, fitted(penreg))
abline(0, 1)
plot(fitted(penreg), residuals(penreg))
```

underprediction of the large value




```
require(lawstat)
runs.test(residuals(penreg))
```

```
    Runs Test - Two sided

data:  residuals(penreg)
Standardized Runs Statistic = -54.131, p-value < 2.2e-16
```

Smoothing splines

```
require(gam)
fitgam <- gam(tobinsQ ~ s(ltdratio) + s(capexratio) + s(rdratio) +
    s(adsratio) + s(pperatio) + s(ebitdaratio) + s(year) + s(assets) +
    s(capex) + s(ltd) + s(ebitda) + s(ppe) + s(sales) + s(ads) +
    s(rd) + s(bookval) + s(mv) + as.factor(indclass), data = newdat)
```

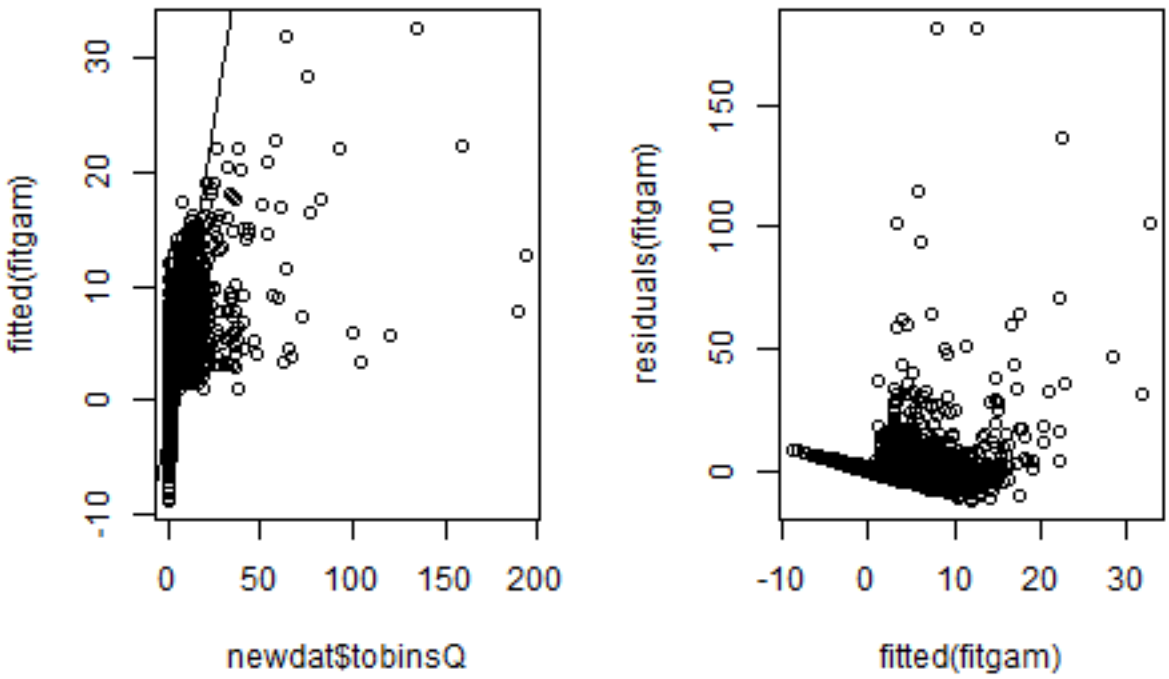
```
summary(fitgam)
```

```
Call: gam(formula = tobinsQ ~ s(ltdratio) + s(capexratio) + s(rdratio) +
    s(adsratio) + s(pperatio) + s(ebitdaratio) + s(year) + s(assets) +
    s(capex) + s(ltd) + s(ebitda) + s(ppe) + s(sales) + s(ads) +
    s(rd) + s(bookval) + s(mv) + as.factor(indclass), data = newdat)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
```

```
-11.6636  -1.3160  -0.4825   0.4117 181.5127

(Dispersion Parameter for gaussian family taken to be 20.593)

    Null Deviance: 375419.6 on 13524 degrees of freedom
Residual Deviance: 276164.2 on 13410.57 degrees of freedom
AIC: 79410.72

Number of Local Scoring Iterations: 6

Anova for Parametric Effects
                      Df Sum Sq Mean Sq   F value    Pr(>F)
s(ltdratio)            1   3555    3555  172.6280 < 2.2e-16 ***
s(capexratio)          1    157     157    7.6472 0.0056940 **
s(rdratio)             1   1446    1446   70.2347 < 2.2e-16 ***
s(adsratio)            1    230     230   11.1753 0.0008312 ***
s(pperatio)            1   1959    1959   95.1115 < 2.2e-16 ***
s(ebitdaratio)         1    640     640   31.0674 2.540e-08 ***
s(year)                1     23      23    1.1060 0.2929696
s(assets)              1    366     366   17.7589 2.524e-05 ***
s(capex)               1    291     291   14.1346 0.0001709 ***
s(ltd)                 1   1208    1208   58.6500 2.012e-14 ***
s(ebitda)              1   9963    9963  483.7935 < 2.2e-16 ***
s(ppe)                 1   2152    2152  104.4790 < 2.2e-16 ***
s(sales)               1   1249    1249   60.6495 7.319e-15 ***
s(ads)                 1    604     604   29.3267 6.220e-08 ***
s(rd)                  1   2426    2426  117.8058 < 2.2e-16 ***
s(bookval)             1   2026    2026   98.4014 < 2.2e-16 ***
s(mv)                  1  60216   60216 2924.0844 < 2.2e-16 ***
as.factor(indclass)   40   4177     104    5.0713 < 2.2e-16 ***
Residuals          13411 276164      21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova for Nonparametric Effects
                   Npar Df Npar F     Pr(F)
(Intercept)
s(ltdratio)             3.0  59.01 < 2.2e-16 ***
s(capexratio)           3.0   2.39 0.0669537 .
s(rdratio)              3.0   5.31 0.0011809 **
s(adsratio)             3.0   9.35 3.589e-06 ***
s(pperatio)             3.0   3.74 0.0106581 *
s(ebitdaratio)          3.0 164.34 < 2.2e-16 ***
s(year)                 3.0   3.14 0.0243951 *
s(assets)               3.0   4.62 0.0030936 **
s(capex)                3.0   3.22 0.0216575 *
s(ltd)                  7.2   7.76 1.462e-09 ***
s(ebitda)               3.0  10.41 7.754e-07 ***
s(ppe)                  4.3  10.21 1.109e-08 ***
s(sales)                3.0   7.48 5.352e-05 ***
s(ads)                  3.0   3.53 0.0141625 *
s(rd)                   3.0   6.66 0.0001722 ***
s(bookval)              3.0 539.86 < 2.2e-16 ***
s(mv)                   3.0 681.05 < 2.2e-16 ***
```

```
as.factor(indclass)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(1, 2))
plot(newdat$tobinsQ, fitted(fitgam))
abline(0, 1)
plot(fitted(fitgam), residuals(fitgam))
```




```
require(lawstat)
runs.test(residuals(fitgam))
```

```
    Runs Test - Two sided

data:  residuals(fitgam)
Standardized Runs Statistic = -60.752, p-value < 2.2e-16
```

Questions

1. TRUE or FALSE? The smoothing parameter for each covariate fitted in the penreg model for the capex and year relationships are very high compared to the other terms. This indicates linearity is likely to result in this case for these terms and we can confirm if linearity is deemed appropriate under this model from the summary output alone (without the need for graphical confirmation).

2. Which of the following is FALSE?

   - The most flexibility afforded to any of the covariates in the penreg model was for bookval which was allocated $df = 9$ for the smoother based term and this is unsurprising when you consider the very small penalty allocated to this term.
   - The year and capex relationships in the penreg model returned large p-values and the pperatio relationship returned a borderline p-value when compared with the 5% level.
   - The residual deviance is approximately 67% of the null deviance for the penreg model.
   - The fitgam model had a smaller percentage of deviance explained compared with the penreg model.
   - The fitgam model also had a smaller estimate for the dispersion parameter than the penreg model - this is very often the case when less variation in the response is explained by the model.

   scale est

3. TRUE or FALSE? The penreg model also underpredicts the largest response values and returns negative fitted values. Including a log link function in this model would not help here (by including a hard boundary at zero) since while including a link function would impose a hard boundary at zero for other spline based models, the extra curvature permitted under a penalised spline-based model could still return negative fitted values (and predictions to new data) even if a log link function was used.

4. TRUE or FALSE? The p-values for each term in the penreg model do not include any uncertainty about the smoothing parameter (estimated using GCV) and this may be important if the uncertainty about the smoothing parameter is large.

5. TRUE or FALSE? One reasonable way to compare the results of the polynomial-based models, B-spline (regression spline based) models, penalised regression spline models and all smoothing spline based models would be to calculate a cross-validation score for each. From these we could use the smallest of these values to choose the best model.

6. Create a multi-choice question with 4 TRUE statements and one FALSE statement with the objective of comparing: regression B-splines, penalised regression splines and smoothing splines.

Be sure to identify the FALSE statement in your answer, and explain why it is false. **This question is worth 5 marks.**

A) B-spline bases are locally defined thus it can better cope will with abrupt changes in the underlying function as long as the knots are sensibly placed.
B) Penalized regression spline usually have a 'large' number of knots.
C) When the flexibility required in the covariate relationship is uneven across the covariate range(s), the fitted curves of both penalized spline and smoothing spline may be sub-optimal.
D) Sometimes the penalized spline and regression spline may return results when smoothingg splines do not as the squared second derivative can be a large burden and models may not converge.
E) Wrong: In smoothing splines, a function which has a smaller amplitude has a 'rougher' curve.
Reason: The function with a bigger amplitude has a 'rougher' curve and the slope is steeper.