Practical 5

Dr L Scott-Hayward

March 2018

Penalized and Unpenalized regression splines for multiple covariates

1. Open your workspace from the first weeks practical.
2. Attach the data set.

Fitting penalized regression splines for multiple covariates

3. Fit a penalized spline based GAM to the data and inspect the results using the following:

```
require(mgcv)
pen_reg<- gam(Nhat ~ s(x.pos)+s(y.pos)+s(Depth)+impact, data=mydata,
family=quasipoisson, offset=log(area))

summary(pen_reg)
```

4. View the partial plots when the:

- y-axes are all on the same scale:

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

- y-axes are all potentially on different scales:

```
par(mfrow=c(2,2))
plot(pen_reg, shade=T, scale=0)
```

5. Visually compare the depth relationship estimated using the `gam` function and a regression spline equivalent with just one knot:

```
require(splines)
par(mfrow=c(1,2))
test<-update(pen_reg, .~.-s(Depth)+bs(Depth, knots=c(20)))
termplot(test, se=T)
```

6. Carry out model selection on the terms in the working GAM using the `dredge` function in the `MuMIn` library. Here is some code to get you started but you'll also need to change the options specification for the treatment of missing values and call the dredge function. It will issue you a warning message but this can be ignored in this case.

```
require(MuMIn)
pen_regPois<-update(pen_reg, family="poisson")
```

```
options(na.action = "na.fail")
dredge(pen_regPois, chat=pen_reg$scale, rank="QAIC")
```

7. Make some predictions for pre and post impact:

- read in the data:

```
preddata<-read.csv("predictionData.csv", header=T)
preddata$Depth<-abs(preddata$Depth)
```

- Use the predict function to make predictions:

```
preds<- predict(pen_reg, newdata=preddata, se=T)
```

- View these predictions pre and post impact:

```
require(fields)
par(mfrow=c(1,2))
quilt.plot(preddata$x.pos[preddata$impact==0],
preddata$y.pos[preddata$impact==0],
preds$fit[preddata$impact==0],
main="Fitted values on the link scale, pre-impact")

quilt.plot(preddata$x.pos[preddata$impact==1],
preddata$y.pos[preddata$impact==1],
preds$fit[preddata$impact==1],
main="Fitted values on the link scale, post-impact")
```

```
par(mfrow=c(1,2))
quilt.plot(preddata$x.pos[preddata$impact==0],
preddata$y.pos[preddata$impact==0],
exp(preds$fit[preddata$impact==0]),
main="Fitted values on the response scale, pre-impact")

quilt.plot(preddata$x.pos[preddata$impact==1],
preddata$y.pos[preddata$impact==1],
exp(preds$fit[preddata$impact==1]),
main="Fitted values on the response scale, post-impact")
```

Fitting an interaction based GAM with multiple covariates

8. Fit a GAM with an interaction between impact and `x.pos`, check out the fitted model and make some predictions based on this new model:

```
pen_regInt<- gam(Nhat ~ s(x.pos, by=impact)+s(y.pos)+s(Depth)+impact,
data=mydata, family=quasipoisson, offset=log(area))
summary(pen_regInt)
preds<- predict(pen_regInt, newdata=preddata, se=T)
```

9. Plot these predictions using the quilt plot code used earlier.
10. Carry out a runs test (from the `lawstat` library) on the Pearsons residuals for this interaction based model.

Questions

1. Which of the following about the pen_reg model results is **TRUE**?
   (a) There is compelling evidence that there are more animals post impact, than pre-impact, on average.
   (b) There is compelling evidence that there are less animals post impact, than pre-impact, on average.
   (c) There is no evidence for any changes in average numbers pre and post impact.
   (d) We cannot tell if there are any differences pre and post impact because the term is not fitted using a smooth function.
   (e) The results may have been different if impact was fitted as using as. factor(impact) rather than as it was fitted here.
2. Which of the following about the pen_reg model results is **FALSE**?
   (a) The F-distribution is the reference distribution for the smoother-based test-statistics because we are estimating the dispersion parameter, if this was assumed to be equal to one the Normal distribution would be used as the reference distribution instead.
   (b) Between 8 and 9 effective degrees of freedom were used to fit each smooth term and the p-values for these smooth terms are small, suggesting these should be retained in the model.
   (c) The percentage of deviance explained by the model reported in the output can be found using: (1-residual deviance/null deviance)*100.
   (d) We could compare various candidate overdispersed Poisson-based GAMs using the QAIC statistic.
   (e) The t-distribution is the reference distribution for the impact-related test-statistic because we are estimating the dispersion parameter, if this was assumed to be equal to one the Normal distribution would be used as the reference distribution instead.
3. TRUE or FALSE? The full model was selected using 'all-possible-subsets' selection governed by the QAIC score.
4. TRUE or FALSE? The second best model based on 'all-possible-subsets' selection omitted impact from the model but the change in the QAIC that occurred by it's omission was relatively small.
5. TRUE or FALSE? The p-values in the pen_reg output associated with each smooth term, compares the fit for models with a smooth term for each covariate and the fit for models with a linear term for each covariate.
6. Which of the following about the partial plots for the pen_reg model is **FALSE**?
   (a) A similar depth relationship can be obtained with just one knot (at 20m) and this results in an improvement in the GCV score compared with the penalized regression based fit.
   (b) All relationships appear to be nonlinear on the link scale, however the depth relationship is less 'wiggly' than the other covariate relationships, despite similar numbers of degrees of freedom being used for all.
   (c) There is a great deal of uncertainty associated with the smooth function for depth beyond 25m.
   (d) While the depth relationship appears to increase again after about 30m, this is within the bounds of uncertainty for the fitted curve and so may not be a genuine increase.
   (e) A similar depth relationship can be obtained with just one knot (at 20m) and this results in better precision (lower uncertainty) in the fitted curve for the larger depths compared with the penalized spline based depth relationship.
7. TRUE or FALSE? Under the interaction based model, there appears to be some redistribution of the animals into the centre of the surveyed area post impact (compared to pre-impact).
8. TRUE or FALSE? The model without interactions is unable to capture the redistribution of the animals into the centre of the surveyed area post impact (compared to pre-impact). This model permits average numbers to change pre and post impact but does not allow the relationship between average numbers and one (or more) of the spatial covariates to change pre and post imapact.

9. TRUE or FALSE? It is possible for the model without interactions to capture the redistribution of the animals into the centre of the surveyed area post impact (compared to pre-impact), because the impact term appears in the model in addition to the spatial covariates.

10. Which of the following is **TRUE**?

    (a) The tests of significance reported by the model output might be incorrect due to the presence of significant positive autocorrelation. In particular, the p-values may be too small.
    (b) The tests of significance reported by the model output are not affected by the presence of significant positive residual autocorrelation, because N is not used in their calculation.
    (c) There is no evidence for the presence of significant positive or negative autocorrelation because the runs test-based test statistic is small.
    (d) The tests of significance reported by the model output might be incorrect due to the presence of significant negative autocorrelation. In particular, the p-values may be too small.
    (e) The tests of significance reported by the model output might be incorrect due to the presence of significant positive autocorrelation. In particular, the p-values may be too large.