DATA 303/473 Test 1 Solutions: Fish markets

1 April 2021

1.

a. **[2 marks]** `speciesPike` had the lowest expected `weight`. Compared to all other species it had the lowest value of $E(\widehat{weight}|species) - E(\widehat{weight}|speciesBream)$, holding all other predictors constant.

b. **[2 marks]** No. None of the numerical predictors had values close to zero; OR the intercept would refer to a fish of zero `vert.len`, `diag.len`, `cross.len`, `height` or `width`, which does not make practical sense.

c. **[2 marks]**

```
weight=-912.7 + 0 -79.84*7.5 + 81.71*28.4 + 30.27*29 + 5.807*7.8 -0.7819*4.2
weight
```

```
## [1] 1728.905
```

d. **[2 marks]** No. The two models can not be compared using BIC as the response variables are measured on different scales.

e. **[3 marks]**

```
exp(0.05541)-1
```

```
## [1] 0.05697389
```

Compared to `speciesBream`, `speciesPike` has a higher expected `weight` by a multiplicative factor of 0.057, holding all other predictors constant.

f. **[3 marks]**

- Smooth terms for `cross.len`, `height` and `width` are non-linear and significant.
- Smooth terms for `vert.len` and `diag.len` are linear and non-significant.

g. **[3 marks]**

There is no evidence that more basis functions are required for any of the smooth terms.

h. **[3 marks]**

AIC preferred model is GAM2. GAM1 and GAM2 have equivalent fit - we use the principle of parsimony and pick the model with fewer predictors.
BIC preferred model is GAM2. It is the model with the lowest BIC value. Model interpretability is less of a concern, so it's not necessary to pick a model with linear terms only. We therefore prefer GAM2 as it is the preferred model according to AIC and BIC and therefore gives the best fit to the data.

2. Write TRUE or FALSE. Where you select FALSE, explain why you think the statement is not true.

a. **[2 marks]** TRUE

b. **[2 marks]** FALSE. Preservation of hierarchy is not required for log transformations.

c. **[2 marks]** FALSE. The model equation given applies to all values of the predictor variables. It is therefore a global basis function.

d. **[2 marks]** FALSE. Model assumptions refer to the population errors. Since residuals are sample estimates of the population errors we should check assumptions based on residuals, not the response variable Y.

e. **[2 marks]** FALSE. This is a linear model since the systematic component $\beta_0 + \beta_1(1/X_1) + \beta_2 \log(X_2) + beta_3 X_3$ is a linear combination of predictors.