DATA 303/473 Test 1: Fish markets

1 April 2021

1. Data were collected on 158 fish of some species commonly sold in fish markets. The variables in the dataset are:

- `species`: Name of fish species
- `weight`: Weight of fish in grams
- `vert.len`: Vertical length in cm
- `diag.len`: Diagonal length in cm
- `cross.len`: Cross length in cm
- `height`: Height in cm
- `width`: Diagonal width in cm

A marine scientist is interested in developing a model that can be used to predict the weight of a fish with `species`, `vert.len`, `diag.len`, `cross.len`, `height` and `width` as potential predictors. Part of the analysis carried out is shown below. Use the results to answer the questions that follow.

```
fish<-read.csv("fishmarket.csv", header=T, stringsAsFactors = TRUE)
summary(fish)
```

```
##       species       weight          vert.len        diag.len
##  Bream    :35   Min.   :   5.9   Min.   : 7.50   Min.   : 8.40
##  Parkki   :11   1st Qu.: 121.2   1st Qu.:19.15   1st Qu.:21.00
##  Perch    :56   Median : 281.5   Median :25.30   Median :27.40
##  Pike     :17   Mean   : 400.8   Mean   :26.29   Mean   :28.47
##  Roach    :19   3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.75
##  Smelt    :14   Max.   :1650.0   Max.   :59.00   Max.   :63.40
##  Whitefish: 6
##    cross.len         height           width
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048
##  1st Qu.:23.20   1st Qu.: 5.941   1st Qu.:3.399
##  Median :29.70   Median : 7.789   Median :4.277
##  Mean   :31.28   Mean   : 8.987   Mean   :4.424
##  3rd Qu.:39.67   3rd Qu.:12.372   3rd Qu.:5.587
##  Max.   :68.00   Max.   :18.957   Max.   :8.142
##
```

```
fit1<-lm(weight~species+ vert.len + diag.len + cross.len +
           height + width, data=fish)
library(pander)
pander(summary(fit1), caption="")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-912.7	127.5	-7.161	3.638e-11
speciesParkki	160.9	75.96	2.119	0.03582
speciesPerch	133.6	120.6	1.107	0.27
speciesPike	-209	135.5	-1.543	0.1251
speciesRoach	104.9	91.46	1.147	0.2532
speciesSmelt	442.2	119.7	3.695	0.0003108

	Estimate	Std. Error	t value	Pr(>|t|)
speciesWhitefish	91.57	96.83	0.9456	0.3459
vert.len	-79.84	36.33	-2.198	0.02955
diag.len	81.71	45.84	1.783	0.07675
cross.len	30.27	29.48	1.027	0.3062
height	5.807	13.09	0.4435	0.6581
width	-0.7819	23.95	-0.03265	0.974

Observations	Residual Std. Error	R^2	Adjusted R^2
158	93.95	0.9358	0.931

```
BIC(fit1)
```

```
## [1] 1937.243
```

```
fit2<-lm(log(weight)~species+ vert.len + diag.len + cross.len +
           height + width, data=fish)
pander(summary(fit2), caption="")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.427	0.2937	8.263	7.889e-14
speciesParkki	0.1541	0.175	0.8803	0.3801
speciesPerch	0.1668	0.2779	0.6002	0.5493
speciesPike	0.05541	0.3122	0.1775	0.8594
speciesRoach	0.1273	0.2108	0.6039	0.5468
speciesSmelt	-1.13	0.2758	-4.097	6.921e-05
speciesWhitefish	0.3183	0.2231	1.427	0.1558
vert.len	0.09876	0.08372	1.18	0.2401
diag.len	-0.1081	0.1056	-1.024	0.3078
cross.len	0.06333	0.06794	0.9321	0.3528
height	0.06754	0.03017	2.239	0.0267
width	0.1973	0.05518	3.575	0.0004756

Observations	Residual Std. Error	R^2	Adjusted R^2
158	0.2165	0.9752	0.9733

```
BIC(fit2)
```

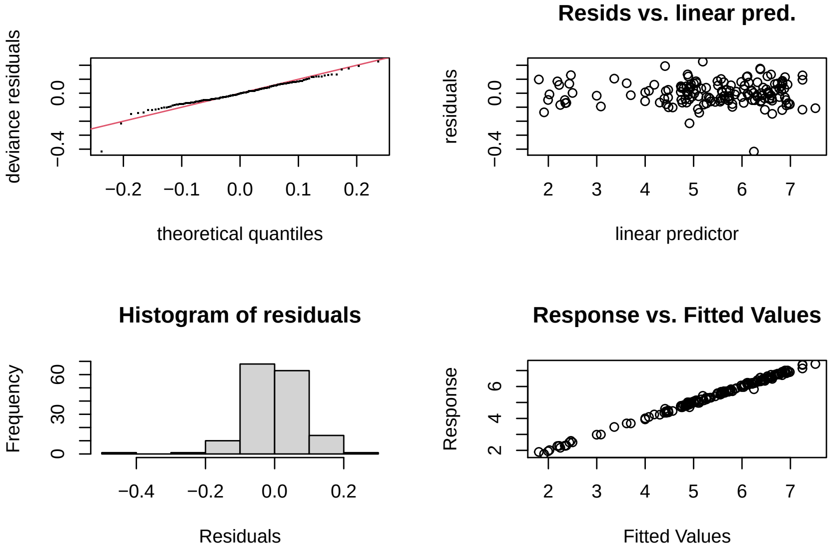
```
## [1] 18.19266
```

```
library(mgcv)
gam1<-gam(log(weight)~species+ s(vert.len) + s(diag.len) + s(cross.len) +
           s(height) + s(width), data=fish, method="REML")
summary(gam1)
```

```
##
## Family: gaussian
## Link function: identity
##
```

```
## Formula:
## log(weight) ~ species + s(vert.len) + s(diag.len) + s(cross.len) +
##     s(height) + s(width)
##
## Parametric coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.32479    0.09485  56.139   <2e-16 ***
## speciesParkki     0.12221    0.08759   1.395   0.1652
## speciesPerch      0.17259    0.13261   1.301   0.1953
## speciesPike       0.11770    0.16907   0.696   0.4875
## speciesRoach      0.10859    0.10506   1.034   0.3032
## speciesSmelt     -0.21700    0.15076  -1.439   0.1524
## speciesWhitefish  0.23760    0.10467   2.270   0.0248 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df      F  p-value
## s(vert.len)  1.000  1.000  1.553 0.214862
## s(diag.len)  1.000  1.000  0.050 0.823955
## s(cross.len) 7.700  8.496 11.347  < 2e-16 ***
## s(height)    3.128  3.999  5.486 0.000404 ***
## s(width)     3.189  4.151  7.133 2.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.996   Deviance explained = 99.6%
## -REML = -123.27  Scale est. = 0.0075377  n = 158
```

```
par(mfrow=c(2,2))
gam.check(gam1)
```



Resids vs. linear pred.

Histogram of residuals

Response vs. Fitted Values

```
##
## Method: REML   Optimizer: outer newton
## full convergence after 10 iterations.
## Gradient range [-4.641604e-05,9.053582e-05]
## (score -123.2722 & scale 0.007537743).
## Hessian positive definite, eigenvalue range [2.131058e-05,73.19175].
## Model rank =  52 / 52
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                 k'  edf k-index p-value
## s(vert.len)  9.00 1.00    0.92    0.14
## s(diag.len)  9.00 1.00    0.97    0.34
## s(cross.len) 9.00 7.70    1.03    0.62
## s(height)    9.00 3.13    1.03    0.60
## s(width)     9.00 3.19    0.92    0.12
```

```r
gam2<-gam(log(weight)~species+ s(cross.len) + s(height) + s(width),
         data=fish,method="REML")
gam3<-gam(log(weight)~species+ cross.len + height + width,
          data=fish, method="REML")
modname<-c("GAM1", "GAM2", "GAM3")
mod.compare<-data.frame(modname,
                        c(AIC(gam1), AIC(gam2),AIC(gam3)),
                        c(BIC(gam1), BIC(gam2), BIC(gam3)))
names(mod.compare)<-c("Model", "AIC", "BIC")
library(pander)
pander(mod.compare,round=3, align='c')
```

Model	AIC	BIC
GAM1	-296.7	-216.9
GAM2	-298	-223.9
GAM3	-24.07	9.615

a. [**2 marks**] Based on the summary output for the model in `fit1` which species had the lowest expected `weight`? Explain your answer briefly.
b. [**2 marks**] Does it make practical sense to interpret the intercept of the model in `fit1`? Explain your answer briefly.
c. [**2 marks**] Using the `fit1` model equation, predict `weight` for a fish with the following characteristics: `species=Bream`, `vert.len`=7.5, `diag.len`= 28.4, `cross.len`=29.0, `height`=7.8, and `width`=4.2.
d. [**2 marks**] *BIC* values for models `fit1` and `fit2` are calculated as shown above. If the analyst said that based on these results, the preferred model is the one with the lower *BIC* value, would you agree? Explain your answer briefly.
e. [**3 marks**] Give a mathematical interpretation of the effect of `speciesPike` on `weight` for the model in `fit2`.
f. [**3 marks**] A GAM is fitted as shown in `gam1`. Comment on the non-linearity and significance of smooth terms.
g. [**3 marks**] Is there evidence that more basis functions are required for any of the smooth terms? Explain your answer briefly.
h. [**3 marks**] *AIC* and *BIC* values are calculated for three models and presented in a table as shown above. State the preferred model according to each criterion. Given that the aim of this modeling exercise is to make accurate predictions of `weight`, which of the three models would you choose as your preferred model? Explain your answer briefly.

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

a. **[2 marks]** Log transformations of the response variable are mainly used to deal with violation of the assumption of normality.

b. **[2 marks]** Preservation of hierarchy is required for polynomial and log transformations.

c. **[2 marks]** The following model equation is an example of a local basis function:

$$Y = \beta_0 + \beta_1(1/X_1) + \beta_2 X_2 + \beta_3 X_2^2 + \epsilon.$$

d. **[2 marks]** In a linear model, checking of model assumptions can be carried out using the response variable Y instead of the residuals.

e. **[2 marks]** The following model is an example of a non-linear model:

$$Y = \beta_0 + \beta_1(1/X_1) + \beta_2 \log(X_2) + \beta_3 X_3 + \epsilon.$$

Test total: 30 marks